خوشه‌یابی تصویر زیرکلمات در متون قدیمی و حجیم چاپی با استفاده از معیار مقایسه تصویری

محمدرضا سهیلی و احسان‌اله کبیر

چکیده: حجم زیاد تصاویر متنی روز به روز مسئله دیجیتالی‌شدن متن تصاویر و همچنین مسئله جستجو در این منابع را اهمیت می‌بخشد. در بازشناسی متن‌های حجیم می‌توان از ویژگی‌هایی مانند محدودبودن تعداد و اندازه قلم، یکسان‌بودن صفحه‌آرایی در کل صفحه‌ها، محدودبودن مجموعه واژه‌ها و حوزه معنایی آنها و یکسان‌بودن سبک نگارشی در کل متن استفاده کرد. در این مقاله الگوریتمی ارائه شده که از یکسان‌بودن نوع و اندازه قلم برای خوشه‌یابی زیرکلمات یک کتاب قدیمی با کیفیت پایین چاپ استفاده شده است. این کتاب ۲۳۳ صفحه دارد و کل زیرکلمات آن که در حدود ۱۱۱۰۰۰ زیرکلمه است جداسازی و برچسب‌زنی شده است. در این تحقیق از یک روش ساده افزایشی برای خوشه‌یابی زیرکلمات استفاده شده است. ابتدا برای هر زیرکلمه چهار ویژگی ساده استخراج می‌شود، در صورتی که تفاوت این ویژگی‌ها از ویژگی‌های نماینده یک خوشه کمتر از مقدار آستانه باشد، مقایسه تصویری بین آن دو انجام می‌شود. به علت زیادبودن تعداد زیرکلمات سعی شده تا از ساده‌ترین روش‌های ممکن استفاده شود تا سرعت اجرا افزایش یابد. نتایج آزمایش‌ها نشان می‌دهد می‌توان زیرکلمات را با دقتی در حدود ۹۹/۷ درصد خوشه‌یابی کرد. نتایج این خوشه‌یابی در مرحله بازشناسی زیرکلمات کمک بسیار زیادی خواهد کرد.

کلید واژه: **تحلیل اسناد تصویری، بازشناسی متون حجیم، خوشه‌یابی افزایشی، جداسازی، مجموعه داده.**

۱- مقدمه

امروزه برای از بین نرفتن اسناد مهم و کتاب‌های قدیمی از بیشتر آنها نسخه دیجیتال تهیه می‌شود و امکان جستجو در متن این منابع از نیازهای مهم به شمار می‌رود. حجم زیاد تصاویر متنی که توسط پروژه‌هایی مانند Google Books [۱] تولید می‌شوند روز به روز مسئله دیجیتالی‌شدن متن تصاویر و همچنین مسئله جستجو در این منابع را اهمیت می‌بخشد. گسترش کتابخانه‌های دیجیتال وابسته به سیستم‌های نمایه‌گذاری و بازیابی دقیق است که محقق نمی‌گردد مگر این که از نرم‌افزارهای OCR قدرتمند برای تولید منابع استفاده شود.

نرم‌افزارهای OCR با کارایی بالا برای زبان انگلیسی وجود دارد و کارایی آنها برای یک صفحه متن، آزمایش شده و مورد قبول است ولی همین نرم‌افزارها برای یک مجموعه مثل یک کتاب، کارایی خوبی ندارند. دقت این نرم‌افزارها برای بازشناسی نویسه در حدود ۹۵ درصد است اما نکته مهمی در این بین است که باید به آن توجه خاص داشت و آن این که خطای ۵ درصد در بازشناسی نویسه، نتیجه قابل قبولی برای حجم زیاد متن نمی‌دهد چرا که اگر به طور متوسط در هر صفحه ۲۰۰۰ نویسه وجود داشته باشد، در یک کتاب ۳۰۰ صفحه‌ای ۳۰۰۰۰ نویسه، غلط شناخته می‌شوند. پیداکردن محل خطاها و تصحیح آنها کار بسیار مشکل و پرهزینه‌ای است.

زمانی که از یک نرم‌افزار OCR معمولی برای بازشناسی متن یک مجموعه بزرگ یا کتاب استفاده می‌شود، خطاهای مشابه در کلیه صفحات تکرار می‌شوند. در یک کتاب، اطلاعاتی غنی برای بازشناسی وجود دارد که می‌تواند دقت بازشناسی حروف و کلمات را به حد چشم‌گیری افزایش دهد. از آنجایی که OCRهای معمولی برای بازشناسی متن در صفحه‌های مجزا طراحی شده‌اند، لذا در طراحی آنها جایی برای یادگیری جهت افزایش کارایی سیستم برای متون حجیم در نظر گرفته نشده است. در حالی که برای افزایش کارایی می‌توان از چند صفحه ابتدای کتاب برای بازشناسی مابقی صفحات آن کمک گرفت.

هدف تحقیق ارائه‌شده در مقاله این است که زیرکلمات یک کتاب فارسی که دارای قلم نامتداول و کیفیت چاپ پایین است، به نحوی خوشه‌یابی شوند که کلمات یکسان درون یک خوشه قرار گیرند. برای این کار از یکسان‌بودن نوع و اندازه قلم در کل متن بهره گرفته شده است. در ادامه می‌توان از نتایج این خوشه‌یابی برای افزایش دقت و یا سرعت یک نرم‌افزار OCR استفاده کرد. در مواردی که کیفیت متن خیلی پایین باشد و یا نرم‌افزار OCR برای آن متن موجود نباشد می‌توان نماینده‌های هر خوشه را به کمک یک کاربر بازشناسی کرد.

در ادامه این مقاله در بخش ۲ مروری بر کارهای انجام‌شده در حوزه بازشناسی متون حجیم انجام شده است. در بخش ۳ مراحل تهیه مجموعه داده و در بخش ۴ الگوریتم پیشنهادی مطرح شده است. در بخش ۵ نتایج آزمایشات ارائه می‌شود و در بخش ۶ این نتایج ارزیابی می‌شوند. بخش ۷ شامل جمع‌بندی و پیشنهادهایی برای ادامه کار است.

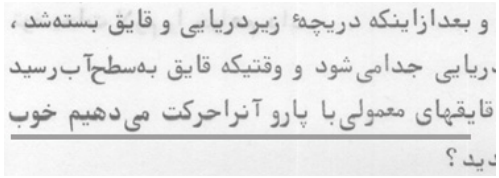
۲- مروری بر کارهای انجام‌شده

تاکنون تحقیقات کمی در زمینه بازشناسی متون حجیم انجام شده و آنها را می‌توان به سه دسته کلی تقسیم کرد که توسط سه گروه پژوهشی انجام شده است. پژوهش‌های گروه اول در راستای انجام یک پروژه بسیار بزرگ بوده و با طراحی معماری یک سیستم شروع شده و در ادامه پژوهش‌هایی در زمینه افزایش دقت، بدون دخالت کاربر و با استفاده از روش‌هایی بر پایه خوشه‌یابی کلمات انجام شده است [۲] تا [۷]. ایده اصلی پژوهش‌های گروه دوم، استفاده از معیار آنتروپی برای تطابق سیستم بازشناسی با متن مورد نظر است [۸] تا [۱۳]. گروه سوم تلاش کرده‌اند تا با ارائه یک روش مبتنی بر شباهت تصویری کلمات، دقت یک نرم‌افزار آماده تجاری را در بخش‌هایی به طور خودکار و در بخش‌هایی با کمک کاربر بهبود بخشند [۱۴].

این مقاله در تاریخ ۱۶ آبان ماه ۱۳۹۱ دریافت و در تاریخ ۴ تیر ماه ۱۳۹۲ بازنگری شد. این پژوهش با استفاده از پشتیبانی مالی مرکز تحقیقات مخابرات ایران بر اساس قرارداد شماره ۵۰۰/۶۹۷۲/ت، کد ۵۳-۰۵-۹۱ انجام شده است.

محمدرضا سهیلی، گروه مهندسی الکترونیک، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، (email: m.soheili@modares.ac.ir).

احسان‌اله کبیر، گروه مهندسی الکترونیک، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، (email: kabir@modares.ac.ir).



شکل ۱: نمونه‌ای از کشیده‌شدن حروف در حین روبش صفحه.

در [۲] راهکاری برای جستجو در یک مجموعه داده شامل ۵۰۰ کتاب اسکن‌شده به زبان تلوگو[۱] ارائه شده است. این مجموعه شامل ۷۵۰۰۰ صفحه بوده و حدود ۲۱ میلیون تصویر کلمه از آن استخراج شده و کلمات کتاب‌های مختلف دارای قلم‌ها، اندازه‌ها و فرمت‌های مختلف بوده است. کل کلمات کتاب به صورت سلسله‌مراتبی خوشه‌یابی شده‌اند، به این شکل که ابتدا با ویژگی‌های ساده خوشه‌یابی زمخت[۲] و سپس با ویژگی‌های پیچیده‌تر خوشه‌یابی ظریف[۳] انجام شده است. ویژگی‌های استفاده‌شده در این تحقیق شامل پروفیل بالایی و پایینی کلمه، پروفیل افکنش[۴] و پروفیل گذار[۵] بوده است. برای هر خوشه یک نماینده انتخاب شده که معیار انتخاب نماینده، کمترین مجموع فاصله از بقیه اعضای خوشه بوده است. برای جستجوی یک کلمه باید از کلمه مورد جستجو تصویر تهیه شود و سپس ویژگی‌های تصویری آن استخراج گردد و در نهایت با ویژگی‌های استخراج‌شده درون درخت ایجادشده در خوشه‌یابی، جستجو انجام گیرد تا خوشه مورد نظر یافت شود. سپس بایستی کلمه مورد جستجو با کلمات داخل خوشه مقایسه گردد.

در [۶] با استفاده از خوشه‌یابی کلمات یک کتاب، دقت بازشناسی افزایش داده شده است. فرض شده کلماتی که درون یک خوشه قرار می‌گیرند، یکسان هستند و با این فرض از نتایج بازشناسی تمام نمونه‌های درون یک خوشه، با استفاده از دو روش CMV[۶] و DTW[۷] برای بازشناسی دقیق‌تر کلمه متناظر با آن خوشه استفاده شده است.

در [۷] کلمات بخشی از کتاب توسط OCR یا کاربر برچسب زده شده و سعی شده تا بازشناسی مابقی کلمات کتاب با استفاده از این کلمات برچسب خورده انجام شود. در این روش ویژگی‌های کلمات استخراج شده و از یک تقریب سریع بر پایه نزدیک‌ترین همسایگی و روش K میانگین سلسله‌مراتبی استفاده می‌شود تا تشخیص دهد که کلمه مورد آزمایش به کدام کلمه در مجموعه یادگیری نزدیک‌تر است.

در [۱۴] کل کلمات کتاب خوشه‌یابی می‌شوند به طوری که کلمات یکسان درون یک خوشه قرار گیرند. هر کلمه‌ای که قرائت شود اگر در خوشه‌های قبلی قرار نگیرد، یک خوشه جدید برای آن ایجاد می‌شود. الگوریتم مقایسه دو کلمه در این تحقیق بر اساس منطبق‌کردن دو کلمه بر روی‌هم[۸] با استفاده از همبستگی متقابل[۹] و شار نوری[۱۰] تغییر یافته و سپس پیداکردن اجزای پیوسته‌ای از اختلاف دو تصویر است که مساحتی بیشتر از چهار نقطه داشته باشند. پس از این که خوشه‌یابی کلمات به طور کامل انجام شد، کلیه کلمات خوشه‌ها به یک OCR تجاری داده شده و سپس از مجموع نتایج خوشه‌یابی و بازشناسی برای افزایش دقت استفاده شده است.

۳- تهیه مجموعه داده

از آنجایی که تاکنون در زمینه متون حجیم فارسی تحقیقی انجام نشده است، مجموعه داده‌ای برای آزمون وجود نداشت بنابراین یکی از کارهایی که باید صورت می‌گرفت تولید یک مجموعه داده برای بررسی آزمایش‌ها و کارایی الگوریتم بود. برای این کار کتاب "بیست هزار فرسنگ زیر دریا"[۱۱] به عنوان مجموعه داده انتخاب شد. این کتاب ۲۳۳ صفحه دارد.

علت انتخاب این کتاب، واحدبودن موضوع کتاب و استفاده از یک نوع قلم در کل کتاب به غیر از عناوین فصل‌ها بوده تا بتوان از این اطلاعات در متن بهره برد. نکته دیگری که مد نظر قرار گرفته است یکسان‌بودن صفحه‌آرایی تمام متن و همچنین نداشتن تصویر در متن می‌باشد تا بتوان مسئله را تا حد امکان ساده کرد و درگیر مشکلات تشخیص نحوه صفحه‌آرایی نشد. در ضمن این کتاب تقریباً ویژگی‌های یک کتاب قدیمی مانند ناآشنا بودن نوع قلم برای نرم‌افزارهای OCR، کیفیت پایین کاغذ و چاپ و کم و زیادشدن جوهر در صفحات مختلف را داراست. همچنین به این علت که در چاپ این کتاب از کاراکتر فاصله با طول مختلف و کشیدن حروف یک کلمه که برای ترازکردن کناره‌های متن به کار می‌رود، استفاده شده است این کتاب تا حدودی ویژگی‌های کتب چاپ جدید را نیز داراست.

۳-۱ روبش صفحات

برای روبش صفحات ابتدا شیرازه کتاب توسط دستگاه بریده شده و کل صفحات کتاب، برگ برگ شده است تا از اعوجاج‌های هندسی ایجادشده در حین روبش تا حد امکان کاسته شود. صفحه‌های کتاب توسط روبشگر HP ۵۵۵۰c دارای کشنده کاغذ، روبش شده است. یکی از مشکلاتی که در حین روبش با این روبشگرها پیش می‌آید کشیده‌شدن عمودی کلمات است. به علت پایین‌بودن کیفیت کاغذ و مستهلک‌شدن چرخ دنده‌های کشنده روبشگر در طول زمان، در بخش‌هایی سرعت حرکت کاغذ در حین روبش کم می‌شود که این باعث کشیده‌شدن کلمات در جهت عمودی می‌شود. نمونه‌ای از این رخداد در شکل ۱ آمده است. به علت این که تشخیص محل وقوع این مشکل، کار دشواری است، در این تحقیق از رفع این اشکال صرف نظر شده است به این امید که در مراحل بعدی خطای زیادی ایجاد نکند.

در مرحله بعد حذف حاشیه‌های کاغذ و همچنین تبدیل کل صفحات به اندازه‌های یکسان انجام شده است. ابتدا ناحیه‌های سیاه اطراف تصویر حذف شده و سپس با استفاده از افکنش افقی و عمودی تصویر، چارچوب اطراف متن تعیین شده است.

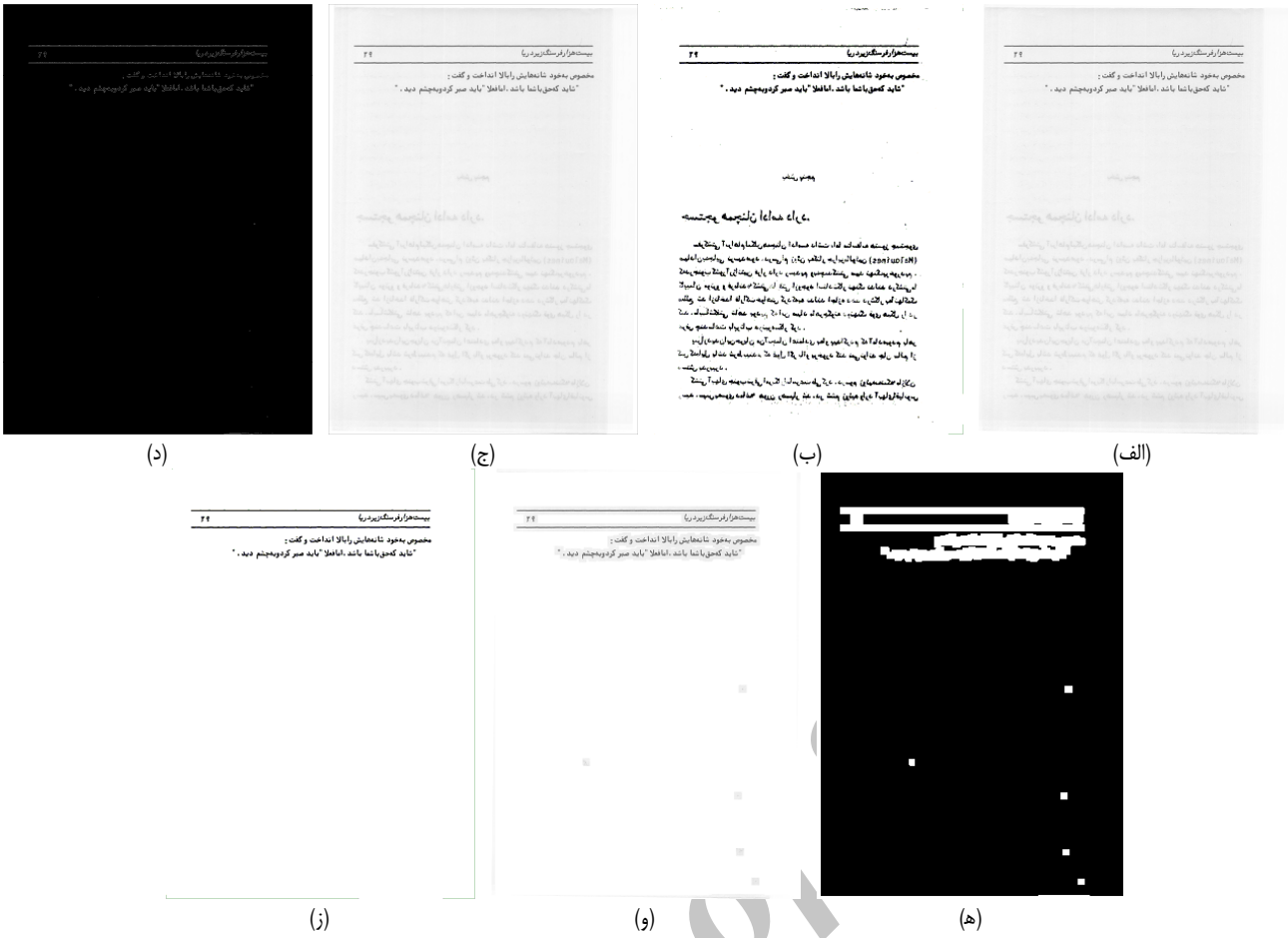
۳-۲ حذف چرخش

با وجود تمهیداتی که در این نوع روبشگرها برای جلوگیری از کجی کاغذ در نظر گرفته شده، باز هم تصاویر به دست آمده از بعضی صفحات چرخش کمتر از ۳ درجه داشت. برای به دست آوردن میزان چرخش

۱. زبان Telugu یک زبان هندی با قدمتی چند هزار ساله است که برای آن نرم‌افزار OCR تجاری وجود ندارد و در حال حاضر ۷۰ میلیون نفر در هند به این زبان صحبت می‌کنند.

2. Coarse
3. Fine
4. Projection Profile
5. Transition Profile
6. Character Majority Voting
7. Dynamic Time Warping
8. Registration
9. Cross Correlation
10. Optical Flow

۱۱. "بیست هزار فرسنگ زیر دریا" اثر ژول ورن، ترجمه امین نصیری، چاپ انتشارات سپیده، چاپ چهارم ۱۳۶۵ (چاپ اول ۱۳۴۶).



(الف) (ب) (ج) (د)

(ه) (و) (ز)

شکل ۲: مراحل پیش‌پردازش و دودویی‌کردن برای یک صفحه از کتاب، (الف) تصویر اولیه، (ب) نتیجه دودویی با استفاده از الگوریتم ولی‌زاده [۱۶]، (ج) تصویر پس از هموارسازی با فیلتر میانگین ۵ در ۵، (د) تصویر پس از عبور از فیلتر سوبل، (ه) ماسک به‌دست آمده از تصویر (د)، (و) تصویر پس از عبور از ماسک و (ز) تصویر دودویی به دست آمده از الگوریتم ولی‌زاده پس از پیش‌پردازش.

تصاویر، ابتدا از تصویر تبدیل فوریه گرفته شده و سپس خطی بر نقاط بیشینه در حوزه تبدیل برازش شده است. با توجه به این که تصاویر، مربوط به متن است لذا در طیف فوریه آنها متناظر با خطوط متن، نقاط بیشینه ظاهر می‌شوند که نشان از پریودیک‌بودن خطوط تصویر دارند. بنابراین خطی که بر این نقاط برازش می‌شود، زاویه خطوط متن را با سطح افق نشان می‌دهد [۱۵].

۳-۳ دودویی‌کردن تصاویر

برای دودویی‌کردن تصاویر از الگوریتم ولی‌زاده [۱۶] استفاده شده است. در این مجموعه متأسفانه در حین روبش تصاویر، متن پشت کاغذ نیز در تصاویر ظاهر شده است که کار دودویی‌کردن را مشکل می‌کند. الگوریتم ولی‌زاده در مورد بیشتر صفحات جواب بسیار خوبی ارائه می‌کند ولی در برخورد با صفحاتی که تعداد خطوط متن آن کمتر از تعداد خطوط متن پشت صفحه است با مشکل مواجه می‌شود. برای حل این مشکل یک پروسه پیش‌پردازش تعریف شده است. از آنجایی که ویژگی اصلی متن‌هایی که از پشت صفحه ظاهر می‌شوند این است که لبه‌های بسیار ضعیفی دارند، ابتدا تصویر با یک فیلتر میانگین ۵ در ۵ هموار شده تا لبه‌های مربوط به متن پشت صفحه کاملاً حذف شود، سپس از لبه‌یاب سوبل برای پیداکردن لبه‌های تصویر استفاده شده است. از نتیجه لبه‌یاب سوبل برای ساختن یک ماسک استفاده شده که این ماسک ناحیه‌هایی از تصویر را که در آن متن وجود ندارد، کاملاً سفید می‌کند. به این ترتیب نقاط مربوط به تصویر پشت کاغذ را در نواحی که نزدیک به نقاط تصویر متن روی کاغذ نیستند، حذف می‌کند. پس از این مرحله از الگوریتم ولی‌زاده استفاده شده که نتایج بسیار خوبی را به دست داده است. تصاویر مربوط به مراحلی که شرح داده شد، برای یک تصویر نمونه در شکل ۲ آمده است.

۳-۴ جداسازی خطوط متن

برای قطعه‌بندی خطوط متن فرض بر این است که خطوط هم‌پوشانی افقی ندارند، بنابراین با استفاده از هیستوگرام افکنش افقی تصویر باید به راحتی بتوان خطوط متن را تشخیص داد. مواردی وجود دارد که باعث می‌شود این کار به سادگی انجام نشود، مانندکم‌بودن فاصله بین خطوط و یکسان‌نبودن ارتفاع خطوط. وجود حروفی مانند آ، ک، م، ع و ح ارتفاع خطوط را زیاد می‌کند و نبود آنها باعث کم‌شدن ارتفاع می‌شود. از موارد دیگر می‌توان به لکه‌های کاغذ و لکه‌های جوهر اشاره کرد. لکه‌های کاغذ معمولاً در کاغذهای کاهی دیده می‌شود و بعضی از این لکه‌ها که ابعادی در حد نقاط متن دارند مشکل‌ساز می‌شوند. تصحیح کلمات چاپی به صورت دست‌نویس که در متون قدیمی رایج بوده است نیز یکی از این موارد است. مورد دیگر یکسان‌نبودن فاصله بین خطوط زمینه و موازی‌نبودن خط‌هاست که در متونی که حروف‌چینی غیر کامپیوتری دارند امری معمولی است. ترازنبودن شروع و پایان خطوط متن نیز یکی دیگر از این مشکلات است. برخی از این مشکلات در شکل ۳ نشان داده شده‌اند.

برای تفکیک خطوط در این تحقیق از هیستوگرام افکنش افقی، نوار زمینه خطوط و فاصله تقریبی بین خطوط استفاده شده است. به علت کم و زیادشدن جوهر در صفحه‌های مختلف، عرض قلم در این صفحات

شکل ۳: نمونه‌هایی از مشکلات موجود در جداسازی خطوط متن، (الف) تصحیح متن به صورت دست‌نویس، (ب) لکه کاغذ، (ج) موازی‌نبودن خطوط و یکسان‌نبودن فاصله بین آنها، (د) لکه جوهر و (ه) ترازنبودن شروع و پایان خطوط متن.

متفاوت است، لذا عرض قلم در هر صفحه جداگانه محاسبه شده است. برای تعیین عرض قلم در هر صفحه ابتدا عرض قلم را در هر خط محاسبه کرده و از بین مقادیر به دست آمده، میانه آن به عنوان عرض قلم انتخاب می‌شود. برای به دست آوردن عرض قلم در هر خط، گذارهای عمودی را برای کل خط به دست آورده و سپس در ستون‌هایی که دو گذار دارند، عرض قلم محاسبه می‌شود. در بین کل مقادیر به دست آمده، میانه آنها به عنوان عرض قلم در آن خط در نظر گرفته می‌شود.

با داشتن عرض قلم و مرز بالا و پایین هر خط، نوار زمینه در آن خط تعیین می‌شود، به این ترتیب که در فاصله بین مرز بالا و پایین هر خط، باید به دنبال ردیف‌های همسایه افقی به تعداد عرض قلم گشت که کمترین مقدار افکنش افقی را داشته باشند.

۳-۵ جداسازی زیرکلمات

مرحله بعد استخراج زیرکلمه است. در خط فارسی به خصوص در متن‌های قدیمی تشخیص دقیق مرز کلمه‌ها تقریباً غیر ممکن است. علت این امر را می‌توان در استفاده‌نکردن از نیم‌فاصله، استفاده نابجا از فاصله و نیم‌فاصله و یا ترازکردن حاشیه متن دانست که فاصله‌های متغیری بین کلمه‌ها ایجاد می‌کند. در متن‌های قدیمی به هم چسبیدگی کلمه‌های مجاور مکرراً مشاهده می‌شود که این امر باعث می‌شود تا قطعه‌بندی با استفاده از اجزای پیوسته نیز به جواب مطلوبی نرسد. لذا در این تحقیق برای جداکردن زیرکلمات از ستون‌های سفید بین زیرکلمات استفاده شده است. نمونه‌هایی از کلمات مجاور به هم چسبیده در شکل ۴ آمده است.

در این تحقیق تعریف در نظر گرفته شده برای زیرکلمه با تعریف رایج آن متفاوت است. بخش‌هایی از یک خط از متن که با ستون‌های سفید از هم جدا شوند به عنوان زیرکلمه در نظر گرفته می‌شود. بنابراین تکه‌هایی

شکل ۴: نمونه‌هایی از کلمات مجاور که به هم چسبیده‌اند.

از کلمات که به هم چسبیده‌اند مانند آنچه در شکل ۴ آمده است به عنوان یک زیرکلمه در نظر گرفته می‌شوند.

۳-۶ برچسب‌زنی زیرکلمات

تعداد زیرکلمات به دست آمده از این کتاب برابر ۱۱۱,۸۱۳ است که به صورت دستی برچسب‌زنی شده‌اند و این کار بسیار سخت با یک کاربر در حدود سه ماه به طول انجامیده است. از این تعداد، ۱۸۹ زیرکلمه به علت خرابی بیش از حد از مجموعه حذف شده‌اند، لذا مجموعه استفاده‌شده در آزمایش‌ها ۱۱۱,۶۲۴ زیرکلمه دارد. برچسب‌های الصاق‌شده به زیرکلمات در مراحل بعدی برای ارزیابی نتایج خوشه‌یابی مورد استفاده قرار می‌گیرند. در مراحل بعدی برای درک بهتر نتایج، به جای برچسب‌زنی دستی از واژه خوشه‌یابی دستی استفاده شده است.

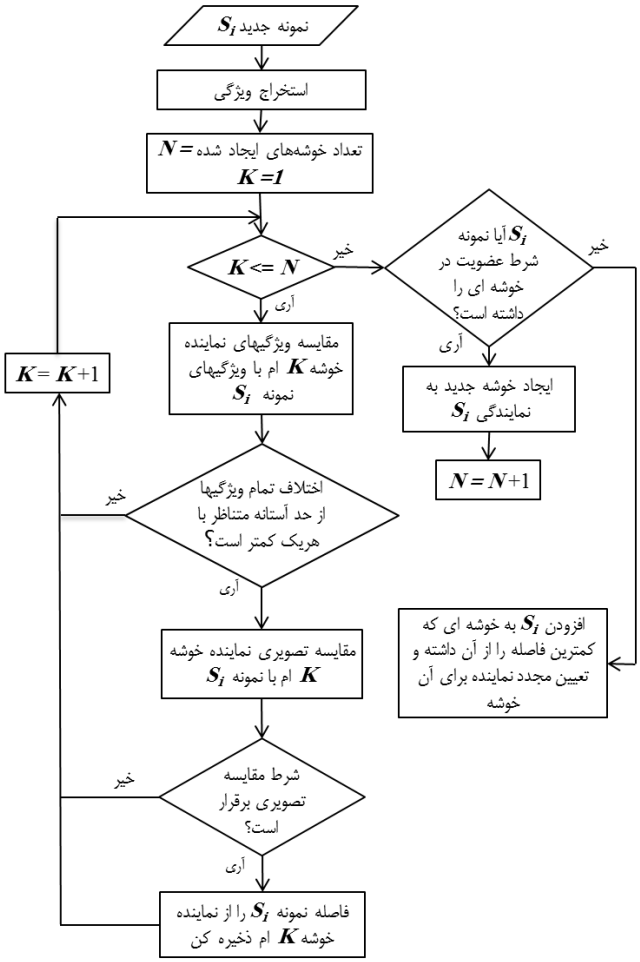
٤- رویکرد پیشنهادی

برای خوشه‌یابی ۱۱۱,۶۲۴ زیرکلمه واضح است که امکان مقایسه تصویری دو به دو وجود ندارد چرا که اگر برای مقایسه هر دو تصویر به ۰/۰۰۳ ثانیه زمان نیاز باشد، مقایسه دو به دوی همه کلمات در حدود ۲۱۶ روز زمان می‌برد. این در حالیست که نویز زیاد متن، کم و زیادشدن جوهر و پارگی و به هم چسبیدگی کلمات، باعث شده تا تفاوت بین زیرکلمه‌های یکسان زیاد باشد و شاید به جرأت می‌توان گفت که برای این مجموعه داده، روشی کارامدتر و دقیق‌تر از مقایسه تصویری نباشد. به همین خاطر در رویکرد پیشنهادی قصد داریم از مقایسه تصویری برای انجام خوشه‌یابی استفاده کنیم. برای خوشه‌یابی زیرکلمات، به علت تعداد زیاد نمونه‌ها و حجم بالای محاسبات، از ساده‌ترین نوع خوشه‌یابی افزایشی استفاده شده است تا بتوان با یک بار پردازش نمونه‌ها و با کمترین محاسبات ممکن، در یک زمان معقول و منطقی عمل خوشه‌یابی را برای یک کتاب به پایان برسانیم. در الگوریتم پیشنهادی سعی شده که از ویژگی‌های استخراج‌شده از زیرکلمات به عنوان نوعی فیلتر و از مقایسه تصویری زیرکلمات به عنوان ابزار تصمیم‌گیری استفاده شود.

٤-۱ الگوریتم خوشه‌یابی زیرکلمات

خوشه‌یابی به این شکل انجام شده است که برای اولین نمونه، یک خوشه ایجاد می‌شود و با مشاهده هر نمونه جدید اگر آن نمونه در یکی از خوشه‌های موجود جای نگرفت یک خوشه جدید برای آن ایجاد می‌شود. برای هر خوشه یک نماینده انتخاب می‌شود که نمونه‌های جدید برای عضویت در آن خوشه، با آن نماینده مقایسه می‌شوند. شکل ۵ روندنمای الگوریتم خوشه‌یابی را نشان می‌دهد.

از هر نمونه جدید ابتدا چهار ویژگی ساده استخراج شده و این ویژگی‌ها با ویژگی‌های نماینده‌های همه خوشه‌های موجود مقایسه می‌شود. مقایسه ویژگی‌ها به این شکل انجام می‌شود که تفاضل ویژگی‌های نمونه مورد نظر از ویژگی‌های متناظر نماینده خوشه محاسبه می‌شود. اگر مقدار تفاضل تمام این ویژگی‌ها از حد آستانه تعریف‌شده برای هر ویژگی کمتر بود، دو نمونه با هم مقایسه تصویری (بخش ۴-۳) می‌شوند. اگر در

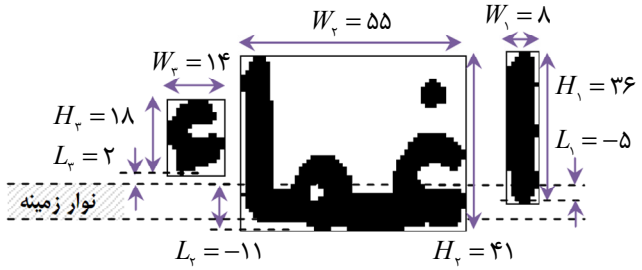


شکل ۵: روندنمای الگوریتم خوشه‌یابی.

مقایسه تصویری، عدم تطابق مشاهده شود نتیجه‌گیری می‌شود که نمونه به خوشه مورد نظر تعلق ندارد، ولی اگر در مقایسه تصویری تطابق زیادی بین دو نمونه دیده شود برای آنها یک فاصله به دست می‌آید که در نهایت در بین خوشه‌ها، خوشه‌ای که کمترین فاصله را داشته باشد به عنوان خوشه مناسب برای این نمونه تعیین می‌شود. اگر در مقایسه یک نمونه با نماینده‌های خوشه‌ها، هیچ یک از نماینده‌ها شرط مربوط به چهار ویژگی را نداشته باشند یا اگر هم شرط را داشته باشند ولی در مقایسه تصویری، تطابق کافی مشاهده نشود، برای آن نمونه، یک خوشه جدید ایجاد می‌شود. ضمناً در صورت عضویت نمونه جدید در یک خوشه، نماینده آن خوشه مجدداً انتخاب می‌شود، به این شکل که نمونه‌ای به عنوان نماینده انتخاب می‌شود که کمترین فاصله را از بقیه نمونه‌های داخل خوشه داشته باشد و این فاصله از راه مقایسه تصویری به دست می‌آید. از آنجایی که در برخی خوشه‌ها مانند خوشه‌های "ا" و "ر" تعداد نمونه‌ها بسیار زیاد است، محاسبه نماینده خوشه به زمان زیادی نیاز دارد لذا در حین خوشه‌یابی فرض شده که اگر خوشه‌ای بیش از ده نمونه داشته باشد، نیازی به انتخاب مجدد نماینده ندارد.

٤-٢ استخراج ویژگی

چهار ویژگی استفاده‌شده شامل عرض، ارتفاع، تعداد نقاط مشکی و سطح شاخص زیرکلمه است. منظور از سطح شاخص زیرکلمه، فاصله پایین زیرکلمه از مرز بالایی نوار زمینه است. اگر قسمت‌هایی از زیرکلمه پایین نوار زمینه باشد سطح شاخص آن منفی و اگر کل زیرکلمه در بالای نوار زمینه باشد سطح شاخص آن مثبت به دست می‌آید. در شکل ۶



شکل ۶: نمایش ویژگی‌های عرض (W)، ارتفاع (H) و سطح شاخص (L) برای زیرکلمه‌های یک کلمه.

سه ویژگی عرض، ارتفاع و سطح شاخص برای زیرکلمه‌های یک کلمه نمایش داده شده است.

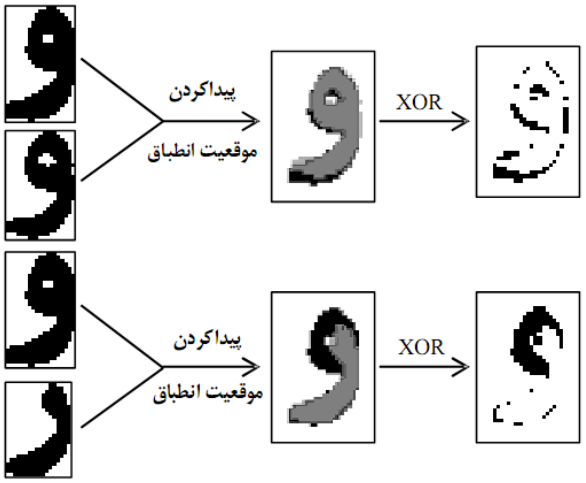
این ویژگی‌ها به علت یکسان‌بودن اندازه و نوع قلم در کل متن، به سادگی می‌توانند تعداد مقایسه‌های تصویری را به شدت کاهش دهند. برای سه ویژگی عرض، ارتفاع و سطح شاخص، مقایسه به این شکل انجام می‌شود که تفاضل مقدار ویژگی نمونه از ویژگی نماینده خوشه باید از یک مقدار آستانه کمتر باشد. این مقدار آستانه برای هر سه ویژگی یکسان در نظر گرفته شده که برابر عرض قلم است. برای ویژگی تعداد نقاط مشکی، مقایسه به این شکل انجام می‌گیرد که ابتدا تعداد نقاط مشکی نمونه و نماینده خوشه از هم کسر می‌شوند که باید مقدار به دست آمده از ۰٫۲۵ کوچک‌ترین آنها کمتر باشد یا به عبارت ساده‌تر در صورتی مقایسه انجام می‌شود که تعداد نقاط مشکی کمتر از ۲۵٪ تفاوت داشته باشد. چون در بسیاری از قسمت‌ها کیفیت متن پایین است، سعی شده تا حد امکان مقادیر آستانه برای ویژگی‌ها بزرگ تعریف شود. با این ترتیب هر جایی که کمترین احتمال یکسان‌بودن زیرکلمه‌ها وجود داشته باشد، مقایسه تصویری انجام می‌شود.

٤-٣ الگوریتم مقایسه تصویری زیرکلمات

در مقایسه تصویری دو زیرکلمه دو نکته مهم هست که باید در مورد آن دقت کافی داشت. اول این که خط فارسی بسیار به نقطه حساس است، لذا در حین مقایسه اگر تفاوت فقط در یک نقطه از یک حرف باشد باز هم دو زیرکلمه نباید در یک خوشه قرار گیرند. دوم این که کم و زیادشدن جوهر در صفحه‌های مختلف باعث ایجاد تفاوت زیاد بین زیرکلمه‌های یکسان می‌شود که گاهی این تفاوت، از اندازه نقطه یک حرف بزرگ‌تر است.

در این تحقیق مقایسه تصویری دو زیرکلمه با استفاده از یک الگوریتم بر پایه تطابق با کلیشه[1] انجام می‌شود، به این ترتیب که ابتدا یک حاشیه سفید به یکی از زیرکلمه‌ها افزوده می‌شود و سپس زیرکلمه دیگر بر روی آن حرکت داده می‌شود تا بهترین موقعیت انطباق به دست آید. بهترین موقعیت، محلی است که تعداد پیکسل‌های روشن در تصویر به دست آمده از XOR دو تصویر در آن حالت کمترین مقدار ممکن را داشته باشد. پس از انطباق دو تصویر، اجزای پیوسته[2] تصویرِ به دست آمده از XOR آنها پردازش می‌شود. در ابتدا اجزای پیوسته‌ای که مساحت آنها کمتر از ۱۶ پیکسل باشد حذف می‌شوند که ۱۶ پیکسل مساحت مربعی به ضلع نصف عرض قلم است. برای مابقی اجزای پیوسته، نسبت مساحت به محیط محاسبه می‌شود و اگر برای یکی از اجزا، این مقدار بزرگ‌تر یا مساوی یک بود، عدم تطابق بین دو زیرکلمه اعلام می‌شود. در غیر این صورت

1. Template Matching
2. Connected Components



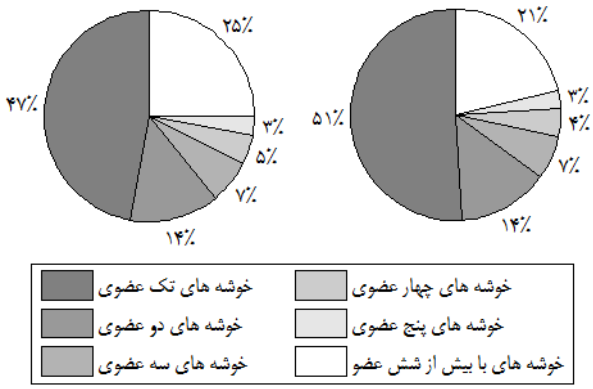
شکل ۷: مراحل مقایسه تصویری دو جفت زیرکلمه.

شکل ۸: برخی از عواملی که باعث شکسته‌شدن خوشه‌ها در خوشه‌یابی خودکار می‌شوند به همراه نمونه‌هایی از زیرکلمات مجموعه داده، (الف) خطای روبشگر، (ب) به هم چسبیدگی حروف، (ج) حذف‌شدن نقاط حروف، (د) پارگی حروف، (ه) کشیدگی حروف، (و) لکه کاغذ و (ز) کم و زیادشدن جوهر.

تعداد پیکسل‌های روشن در تصویر به دست آمده از XOR دو زیرکلمه به عنوان فاصله دو زیرکلمه منظور می‌شود. عدد یک انتخاب‌شده به عنوان مقدار آستانه، از تقسیم مساحت بر محیط دایره‌ای به قطر نصف عرض قلم به دست آمده است. در شکل ۷ روال مقایسه تصویری زیرکلمات برای دو جفت زیرکلمه نمایش داده شده است. در مقایسه دو حرف "و" نسبت مساحت به محیط برای بزرگ‌ترین جزء پیوسته ۰/۴۶ بوده در حالی که در مقایسه حرف "و" و حرف "ر" این مقدار ۱/۹۲ است. در مقایسه دو حرف "و" چون شرط مقایسه تصویری برقرار شده است فاصله دو نمونه محاسبه می‌شود که در این مثال ۵۷ است.

۵- نتایج خوشه‌یابی

در خوشه‌یابی دستی، زیرکلمات در ۵۱۵۴ خوشه جای گرفته‌اند. این در حالیست که در خوشه‌یابی خودکار ۷۸۷۷ خوشه تشکیل شده است. از ۲۷۲۳ خوشه اضافی، ۴۲۲ خوشه به علت کشیدگی حروف درون کلمات برای ترازکردن حاشیه متن بوده است. کم و زیادشدن جوهر، پارگی و به هم چسبیدگی حروف، لکه‌های جوهر و کاغذ، تغییر شکل حروف در اثر خطای روبشگر و حذف‌شدن نقاط حروف در اثر خطای دودویی‌کردن از جمله موارد دیگری هستند که باعث افزایش تعداد خوشه‌ها می‌شوند. تعدادی از نمونه‌هایی که دارای این مشکلات هستند در شکل ۸ نمایش داده شده است.



شکل ۹: نمودار درصد خوشه‌ها بر حسب تعداد اعضای خوشه برای خوشه‌یابی دستی (چپ) و خودکار (راست).

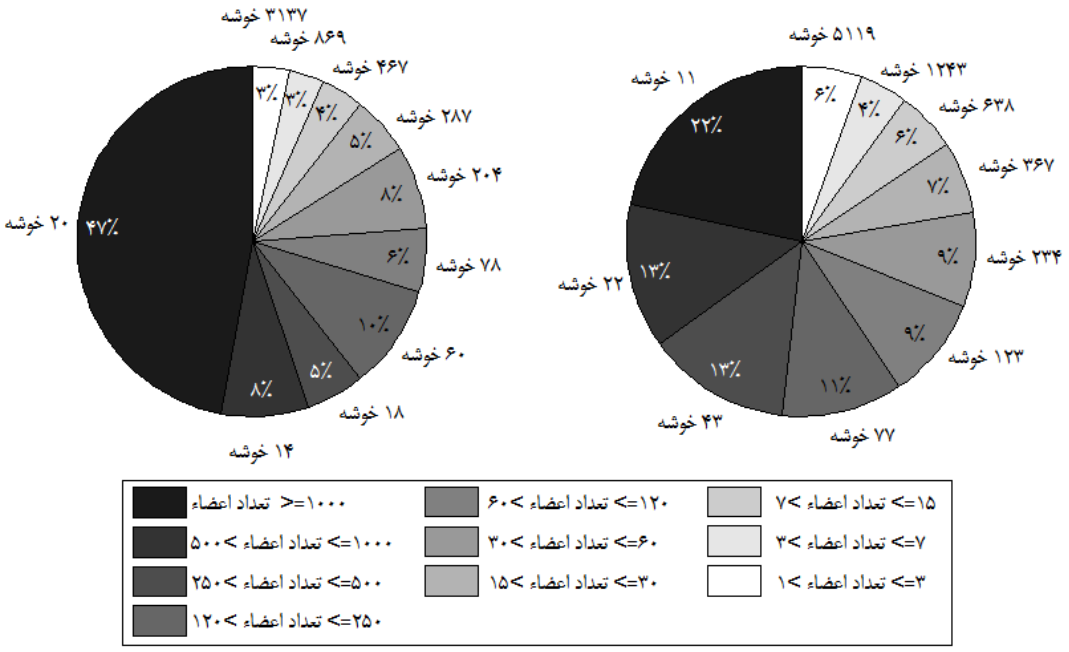
جدول ۱: خوشه‌های با بیش از ۱۰۰۰ عضو و تعداد اعضای آنها در هر دو نوع خوشه‌یابی دستی (۲۰ خوشه) و خودکار (۱۱ خوشه).

خوشه‌یابی خودکار				خوشه‌یابی دستی	
خوشه	تعداد اعضا	خوشه	تعداد اعضا	خوشه	تعداد اعضا
ا	۹۴۶۷	به	۱۶۷۷	د	۶۲۷۱
د	۸۹۶۸	که	۱۶۴۲	ا	۴۳۶۴
ر	۳۴۲۵	با	۱۵۱۸	ا	۲۲۰۹
،	۲۹۴۸	ما	۱۴۲۸	د	۲۱۷۵
و	۲۹۲۱	آ	۱۳۷۶	ر	۱۸۰۶
.	۲۷۰۵	م	۱۳۴۱	ه	۱۳۵۱
ی	۲۲۶۶	ند	۱۲۳۶	ر	۱۳۱۸
ن	۲۲۳۸	ین	۱۱۲۵	ی	۱۳۰۵
را	۱۸۴۰	نا	۱۰۵۱	و	۱۲۷۲
ز	۱۷۶۷			ا	۱۰۲۷
ه	۱۷۱۰			ن	۱۰۰۷

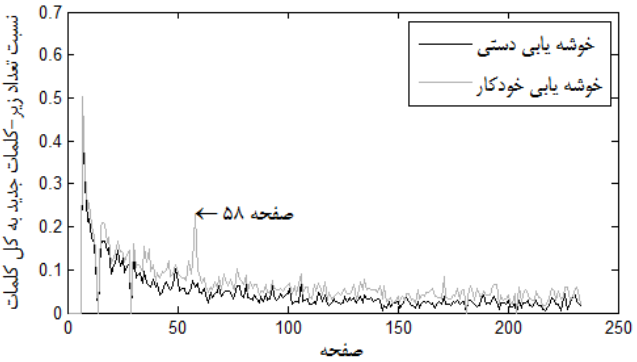
در خوشه‌یابی دستی، در حدود ۴۷٪ خوشه‌ها فقط حاوی یک نمونه هستند که این مقدار در خوشه‌یابی خودکار به حدود ۵۱٪ افزایش یافته است. نمودار درصد خوشه‌ها بر حسب تعداد اعضای خوشه برای خوشه‌یابی دستی و خودکار در شکل ۹ نمایش داده شده است. همان طور که ملاحظه می‌شود، توزیع نمونه‌ها در خوشه‌ها در خوشه‌یابی خودکار نسبت به خوشه‌یابی دستی تغییرات اندکی دارد.

شکل ۱۰ از زاویه‌ای دیگر حقیقت داده‌ها را نشان می‌دهد. در این نمودار خوشه‌ها بر حسب تعداد اعضا به ده دسته تقسیم شده‌اند و درصد نمونه‌هایی که در هر دسته قرار گرفته، در شکل آمده است. در خوشه‌یابی دستی تنها ۲۰ خوشه، حدود نیمی از نمونه‌ها را در خود جای داده‌اند. این در حالیست که در خوشه‌یابی خودکار، بسیاری از خوشه‌ها به عللی که برخی از آنها در شکل ۸ آمده است به چندین خوشه شکسته می‌شوند.

بدیهی است که شکسته‌شدن خوشه‌ها باعث کاهش تعداد اعضای خوشه‌ها و پراکنده‌شدن نمونه‌ها می‌شود. در خوشه‌یابی دستی، خوشه‌های مربوط به حروف "ا"، "د" و "ر" به ترتیب با ۹۴۶۷، ۸۹۶۸ و ۳۴۲۵ عضو، بیشترین تعداد نمونه را در خود جای داده‌اند و در خوشه‌یابی خودکار خوشه‌های مربوط به حروف "د"، "ا"، "ا"، "د" و "ر" به ترتیب با ۶۲۷۱، ۴۳۶۴، ۲۲۰۹، ۲۱۷۵ و ۱۸۰۶ عضو، بیشترین تعداد نمونه را در خود جای داده‌اند. علت پیشی‌گرفتن حرف "ا" از حرف "د" در خوشه‌یابی خودکار، نازک‌بودن حرف "ا" و آسیب‌پذیری بیشتر آن نسبت به حرف "د" است. جدول ۱ خوشه‌های با بیش از ۱۰۰۰ عضو را هر دو نوع خوشه‌یابی دستی



شکل ۱۰: نمودار توزیع نمونه‌ها در بین خوشه‌ها بر حسب تعداد نمونه‌های هر خوشه برای خوشه‌یابی دستی (چپ) و خودکار (راست).



شکل ۱۱: نمودار نسبت زیرکلمات جدید به کل کلمات در هر صفحه.

جدول ۲: تعداد و نوع خطای اتفاق‌افتاده در خوشه‌یابی خودکار.

اختلاف	تعداد	اختلاف	تعداد	اختلاف	تعداد
یک نقطه	۲۲۸	نقطه و کاما	۷	"ه" و صفر	۱
دو نقطه	۱۲	"ر" و "و"	۶	کلاه الف	۱
سه نقطه	۲۴	"ع" و "ه"	۴	"د" و "ه"	۱
چهار نقطه	۲	"ب" و "ه"	۳	"ی" و "و"	۱
"ک" و "گ"	۴۳	"ا" و "!"	۳	"ا" و "ا"	۱
"س" و "ـ"	۷	"ه" و "ه"	۲		

جدول ۳: یک نمونه صحیح و یک نمونه خطا از چند خوشه دارای خطا.

صحیح	خطا	صحیح	خطا
ا	!	نگ	نک
آ	آ	شما	شعا
ه	ه	مرا	ەرا
فه	فد	شست	شـت

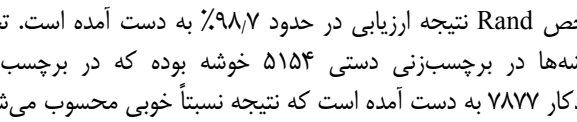
و خودکار نشان می‌دهد. از ۱۱۱۶۲۴ زیرکلمه فقط ۳۴۲ کلمه در خوشه اشتباه قرار گرفته‌اند. ۲۲۸ زیرکلمه (تقریباً ۶۷٪) فقط به علت تفاوت در یک نقطه و ۴۳ نمونه (تقریباً ۱۳٪) به علت پخش‌شدگی جوهر و عدم تشخیص "ک" از "گ" به اشتباه خوشه‌یابی شده‌اند. تمام خطاها به همراه علت وقوع آنها در جدول ۲ آمده است.

در بیش از نیمی از این خطاها نقطه‌های کلمات به قدری محو و کوچک شده‌اند که تشخیص آن بر اساس الگوهای تصویری کار دشواری است. در پس‌پردازش به کمک یک واژه‌نامه بخشی از این خطاها تصحیح می‌شوند. جدول ۳ یک نمونه خطا و یک نمونه صحیح از برخی خوشه‌ها که دچار خطا شده‌اند را نشان می‌دهد.

شکل ۱۱ نسبت تعداد زیرکلمات جدید به کل کلمات را در هر صفحه در هر دو روش خوشه‌یابی دستی و خودکار نشان می‌دهد. در صفحه ابتدایی حدود ۵۰٪ زیرکلمات یکتا بوده‌اند و برای آنها خوشه تشکیل شده است. به مرور در صفحات بعدی تعداد زیرکلمات جدید کاهش پیدا کرده است به طوری که پس از پردازش حدود بیست صفحه از کتاب، تعداد زیرکلمات جدید در هر صفحه به کمتر از ۱۵٪ و پس از پردازش حدود چهل صفحه به کمتر از ۱۰٪ کاهش می‌یابد.

نکته دیگری که در نمودار شکل ۱۱ قابل مشاهده است یکسان‌بودن رفتار هر دو نوع خوشه‌یابی در ساخت خوشه‌های جدید است. بالاتربودن پوش نمودار خوشه‌یابی خودکار نسبت به نمودار خوشه‌یابی دستی به علت تشکیل خوشه‌های اضافی است که در ابتدای این بخش در مورد آن بحث شد. تنها تفاوت شدیدی که در دو نمودار دیده می‌شود مربوط به قله‌ای است که برای صفحه ۵۸ تشکیل شده است. ارتفاع نمودار در این نقطه نسبت به روند کاهشی که در نمودار خوشه‌یابی خودکار دیده می‌شود نیز غیر منتظره است. پس از بررسی انجام‌شده معلوم شد که پیدایش این قله به این علت است که کیفیت چاپ در صفحه ۵۸ نسبت به صفحه‌های دیگر تفاوت شدیدی دارد. به علت کم‌شدن شدید جوهر در این صفحه و عدم جبران آن در بخش دودویی‌کردن، بخش مقایسه تصویری به ازای بسیاری از کلمه‌هایی که قبلاً مشاهده شده بودند، خوشه جدید ایجاد کرده است. بخشی از تصویر این صفحه در شکل ۱۲ آمده است.

نکته مهمی که از تشکیل این قله می‌توان درک کرد این است که وجود قله‌های نامتعارف در این نمودار می‌تواند به نوعی نشان از وقوع خرابی در یکی از مراحل پیش‌پردازش یا خوشه‌یابی در صفحه متناظر با



شکل ۱۲: بخشی از صفحه ۵۸ کتاب که به علت کم‌شدن جوهر باعث تولید خوشه‌های جدید شده است.

آن قله باشد و می‌توان از آن به عنوان پس‌خورد مناسبی جهت بهبود عملکرد سیستم استفاده کرد.

خوشه‌یابی توسط یک رایانه با سیستم عامل ۶۴ بیتی Windows ۷، با ۱۶ گیگابایت RAM و پردازنده سری i۷ توسط نرم‌افزار Matlab ۲۰۱۲ انجام شده است. خوشه‌یابی زیرکلمات توسط این رایانه در حدود ۵۴ ساعت طول کشید که نسبتاً زمان زیادی است و می‌توان با بهینه‌سازی کدهای برنامه و تبدیل کد برنامه به کدهای هم‌روال این زمان را کاهش داد.

۶- ارزیابی نتایج

چون نتیجه مطلوب خوشه‌یابی در این مسئله مشخص است، برای ارزیابی نتایج خوشه‌یابی می‌توان از معیارهایی مانند خلوص[1] و یا شاخص Rand[2] استفاده کرد [۱۷]. برای محاسبه معیار خلوص فرض بر این است که پس از انجام خوشه‌یابی، در هر خوشه هر زیرکلمه‌ای که بیشترین فراوانی را داشته باشد به عنوان زیرکلمه صحیح و مابقی زیرکلمه‌های داخل خوشه به عنوان زیرکلمه‌های نادرست در نظر گرفته شوند. از تقسیم مجموع تعداد زیرکلمه‌هایی که در تمام خوشه‌ها به صورت صحیح خوشه‌یابی شده‌اند بر تعداد کل زیرکلمه‌های مجموعه، مقدار خلوص محاسبه می‌شود. در این آزمایش مقدار خلوص برابر ۰/۹۹۶۹ به دست آمده که نشان از وقوع حدود ۰/۳٪ خطا دارد. البته معیار خلوص برای ارزیابی مشکلی دارد، به این ترتیب که هرچه تعداد خوشه‌ها افزایش پیدا کند از تعداد خطاها کاسته شده و در نتیجه معیار خلوص به سمت ۱ میل می‌کند که این مطلوب نیست.

شاخص Rand از (۱) محاسبه می‌شود [۱۷]

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \qquad (۱)$$

که مقادیر TP[3]، TN[4]، FP[5] و FN[6] با این فرض محاسبه شده که خوشه، متناظر با نمونه‌هایی است که تعداد آنها در آن خوشه در اکثریت باشد. در این آزمایش مقادیر TP، TN، FP و FN به ترتیب ۴۹۸۹۱۱۰۶، $۶/۱ \times ۱۰^{۹}$، ۲۹۴۲۵ و ۷۷۳۳۶۰۴۱ به دست آمده و مقدار شاخص Rand ۰/۹۸۷۶ شده است.

۷- جمع‌بندی

در این مقاله الگوریتمی برای خوشه‌یابی زیرکلمات یک کتاب ارائه شده و نتایج آزمایش آن بررسی شده است. نتایج نشان می‌دهد که خوشه‌یابی با معیار خلوصی در حدود ۹۹/۷٪ انجام شده است. همچنین با معیار شاخص Rand نتیجه ارزیابی در حدود ۹۸/۷٪ به دست آمده است. تعداد خوشه‌ها در برچسب‌زنی دستی ۵۱۵۴ خوشه بوده که در برچسب‌زنی خودکار ۷۸۷۷ به دست آمده است که نتیجه نسبتاً خوبی محسوب می‌شود. همچنین برای آزمودن الگوریتم، مجموعه داده بسیار خوبی ایجاد شده که برای انجام تحقیقات بعدی می‌تواند راهگشا باشد.

برای ادامه کار به نظر می‌رسد اضافه‌کردن یک یا دو کتاب دیگر به مجموعه داده بتواند بحث بر روی کارایی الگوریتم را دقیق‌تر کند. همچنین بررسی تأثیر افزودن ویژگی‌هایی به ویژگی‌های استخراج‌شده که مقاومت زیادی نسبت به نویز داشته باشند می‌تواند در تحقیقات آینده بسیار سودمند باشد.

در تحقیقات آینده می‌توان تعداد خوشه‌های تشکیل‌شده در هر صفحه را به عنوان یک بازخورد در مرحله میانی یک نرم‌افزار OCR مورد بررسی قرار داد.

برای کاهش خطا پیشنهاد می‌شود که یک روال پس‌پردازش مجدداً تمام نمونه‌ها را مورد آزمون قرار دهد و تعلق نمونه به خوشه یا خوشه‌های نزدیک به آن را بررسی کند. برای این کار می‌توان نوعی رابطه یا همسایگی بین خوشه‌ها تعریف کرد. برای تعریف این همسایگی می‌توان از ویژگی طول و عرض نماینده خوشه و فاصله به دست آمده از مقایسه تصویری دو به دوی نماینده‌های خوشه‌ها استفاده کرد. همچنین پیشنهاد می‌شود که در مرحله اول خوشه‌یابی برای هر نمونه اولویت‌های دوم و سوم نیز مشخص شود تا بتوان در مرحله پس‌پردازش از آن استفاده کرد.

مراجع

[1] http://en.wikipedia.org/wiki/Google_Books

[2] K. Pramod Sankar and C. V. Jawahar, "Enabling search over large collections of telugu document images - an automatic annotation based approach," in *Proc. of the 5th Indian Conf. on Computer Vision, Graphics, and Image Processing, ICVGIP*, vol. 4338, pp. 837-848, Dec. 2006.

[3] K. Pramod Sankar, V. Ambati, L. Pratha, and C. V. Jawahar, "Digitizing a million books: challenges for document analysis," in *Proc. of the 7th IAPR Int. Workshop on Document Analysis Systems, DAS'06*, vol. 3872, pp. 425-436, Feb. 2006.

[4] M. Meshesha and C. V. Jawahar, "Self adaptable recognizer for document image collections," in *Proc. of the 2nd Int. Conf. on Pattern Recognition and Machine Intelligence*, vol. 4815, pp. 560-567, Dec. 2007.

[5] N. V. Neeba and C. V. Jawahar, "Recognition of books by verification and retraining," in *Proc. of the 19th Int. Conf. on Pattern Recognition, ICPR'08*, 4 pp., Dec. 2008.

[6] V. Rasagna, A. Kumar, C. V. Jawahar, and R. Manmatha, "Robust recognition of documents by fusing results of word clusters," in *Proc. of the 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09*, pp.566-570, Jul. 2009.

[7] K. Pramod Sankar, C. V. Jawahar, and R. Manmatha, "Nearest neighbor based collection OCR," in *Proc. of the 9th IAPR International Workshop on Document Analysis Systems, DAS'10*, pp. 207-214, 2010.

[8] P. Xiu and H. S. Baird, "Whole-book recognition using mutual-entropy-driven model adaptation," in *Proc. 15th Document Recognition and Retrieval Conf., DRR'08*, vol. 6815, 2008.

[9] P. Xiu and H. S. Baird, "Towards whole - book recognition," in *Proc. of the 8th IAPR Int. Workshop on Document Analysis Systems, DAS'08*, pp.629-636, Sep. 2008.

[10] P. Xiu and H. S. Baird, "Scaling up whole-book recognition," in *Proc. of the 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09*, pp.698-702, Jul. 2009.

[11] P. Xiu and H. S. Baird, "Analysis of whole-book recognition," in *Proc. of the 9th IAPR Int. Workshop on Document Analysis Systems, DAS'10*, pp. 199-206, 2010.

[12] P. Xiu and H. S. Baird, "Incorporating linguistic post-processing into whole-book recognition," in *Proc. of the 17th Document Recognition and Retrieval Conf., DRR'10*, Jan. 2010.

1. Purity
2. Rand Index
3. True Positive
4. True Negative
5. False Positive
6. False Negative

[13] P. Xiu and H. S. Baird, "Incorporating linguistic model adaptation into whole-book recognition," in *Proc. of the IAPR 20th Int. Conf. on Pattern Recognition, ICPR'10*, pp.2057-2060, Aug. 2010.
[14] V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos, "Word-based adaptive OCR for historical books," in *Proc. of the 10th Int Conf. on Document Analysis and Recognition, ICDAR'09*, pp.501-505, Jul. 2009.
[15] J. J. Hull, "Document image skew detection: survey and annotated bibliography," *Document Analysis Systems II, World Scientific*, pp. 40-64, 1998.
[16] M. Valizadeh and E. Kabir, "Binarization of degraded document image based on feature space partitioning and classification," *Int. J. on Document Analysis and Recognition*, vol. 15, no. 1, pp. 57-69, 2012.
[17] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

محمدرضا سهیلی تحصیلات خود را در مقاطع کارشناسی مهندسی کامپیوتر در سال ۱۳۷۸ از دانشگاه تهران و در مقاطع کارشناسی ارشد معماری کامپیوتر در سال ۱۳۸۰ از دانشگاه تربیت مدرس به پایان رسانده است و از سال ۱۳۸۹ تاکنون دانشجوی دکتری الکترونیک در دانشگاه تربیت مدرس می‌باشد. نام‌برده از سال ۱۳۸۱ به عنوان عضو هیئت علمی دانشگاه خوارزمی مشغول به کار می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش و بازشناسی اسناد تصویری، پردازش تصویر، بینایی ماشین و کاربردهای آن و نهان‌نگاری تصاویر دیجیتال.

احسان‌اله کبیر در دهم آبان ۱۳۳۷ در تهران به دنیا آمد. او کارشناسی ارشد پیوسته خود را در مهندسی برق و الکترونیک از دانشکده فنی دانشگاه تهران و دکترای خود را در مهندسی سیستم‌های الکترونیک از دانشگاه اسکس در انگلستان، به ترتیب در سال‌های ۱۳۶۴ و ۱۳۶۹ دریافت کرد. او اکنون استاد دانشکده مهندسی برق و کامپیوتر دانشگاه تربیت مدرس است. زمینه پژوهشی مورد علاقه او بازشناسی الگو، به ویژه بازشناسی متون چاپی و دستنویس است.