استفاده از مناطق شاخص زیر-کلمات چاپی فارسی برای کاهش فضای جستجو در بازشناسی آنها

هما داودی و احسان‌اله کبیر

چکیده: در روش‌های رایج برای کاهش اندازه دیکشنری، معمولاً مجموعه کلمات بر اساس ویژگی‌های شکل کلی‌شان خوشه‌بندی می‌شوند. سپس، هر کلمه ورودی به این خوشه‌ها طبقه‌بندی می‌شود. با توجه به تأثیر مستقیم این مرحله در نتیجه نهایی سیستم بازشناسی، کاهش دیکشنری باید با دقت بالایی انجام شود. به این منظور در این مقاله روشی را برای تأیید ارائه می‌کنیم که میزان اطمینان به خوشه انتخابی را تعیین می‌کند. میزان اطمینان به خوشه انتخابی بر اساس ویژگی‌های محلی شکل تعیین می‌شود. بردارهای ویژگی محلی از شکل زیر- کلمه ورودی استخراج شده و با مناطق شاخص متناظر با خوشه انتخابی مقایسه می‌شود. مناطق شاخص یک خوشه، مناطقی از شکل هستند که زیر- کلمات آن خوشه را از سایر خوشه ها متمایز می‌کنند و در انتها روش تأیید پیشنهادی به همراه مجموعه‌ای از قوانین برای کاهش اندازه دیکشنری به کار می‌رود. آزمایش‌های انجام‌شده بر مجموعه شکل‌های زیر- کلمات فارسی نشان می‌دهد با روش پیشنهادی این مقاله می‌توان با حفظ دقت، فضای جستجو را تا حد قابل توجهی کاهش داد.

کلید واژه: **تأیید، توصیف‌گر شکل، زیر- کلمات چاپی، شکل کلمات، طبقه‌بندی، منطقه شاخص.**

۱- مقدمه

روش‌های بازشناسی کلمات به ۲ دسته کلی تقسیم می‌شوند: روش‌های مبتنی بر قطعه‌بندی و روش‌های مبتنی بر شکل کلی. در روش‌های مبتنی بر قطعه‌بندی، تصویر هر کلمه ابتدا به دنباله‌ای از زیر- تصاویر تجزیه می‌شود که سعی می‌شود هر کدام، نشان‌دهنده یک حرف باشد. با ترکیب نتایج حاصل از بازشناسی این زیر- تصاویر، کل کلمه بازشناسی می‌شود. در روش‌های مبتنی بر شکل کلی، توصیف هر کلمه بر اساس ویژگی‌های کلی شکل آن انجام می‌شود. بر خلاف روش مبتنی بر قطعه‌بندی در این روش هر کلمه یک‌پارچه در نظر گرفته شده و بازشناسی در سطح کلمه انجام می‌شود. در روش‌های بازشناسی مبتنی بر شکل کلی برای هر کلمه یک کلاس در نظر گرفته می‌شود. به این ترتیب، ویژگی‌های منحصر به شکل هر کلمه نیز در بازشناسی وارد می‌شوند. کارایی روش‌های مبتنی بر توصیف شکل کلی در سامانه‌های بازشناسی کلمات در تحقیقات مختلف آمده است [۱] تا [۴]. علاوه بر سامانه‌های بازشناسی، استفاده از اطلاعات شکل کلی کلمه در بازیابی میان مجموعه محدود کلمات، حجم پردازش را به شکل قابل ملاحظه‌ای کاهش می‌دهد. همچنین، توصیف شکل کلی کلمه، روشی کارامد برای نشان‌کردن کلمات پرس و جو در تصاویر اسناد است [۵] تا [۱۱].

این مقاله در تاریخ ۲۹ دی ماه ۱۳۹۱ دریافت و در تاریخ ۱۴ شهریور ماه ۱۳۹۲ بازنگری شد.

هما داودی، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، (email: h.davoudi@modares.ac.ir).

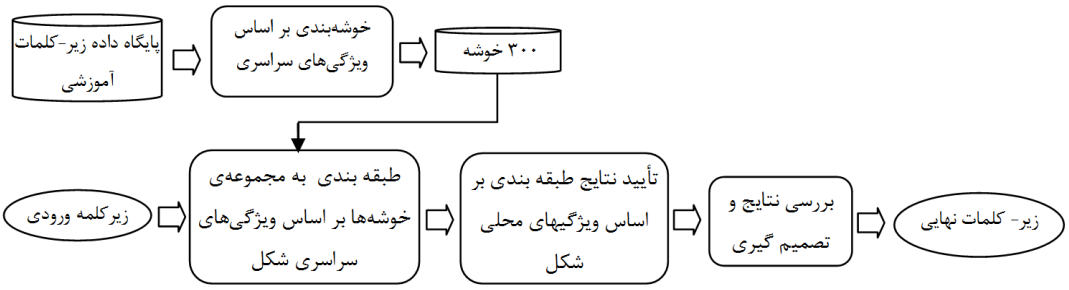
احسان‌اله کبیر، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، (email: kabir@modares.ac.ir).

در بازشناسی کلمات بر اساس شکل کلی، هر کلمه یک کلاس منحصر به فرد را تشکیل می‌دهد. با توجه به تعداد زیاد زیر- کلمات رایج در زبان فارسی (حدود ۱۳۰۰۰ زیر- کلمه)، بررسی این تعداد کلاس از اصلی‌ترین چالش‌های موجود در روش‌های مبتنی بر شکل کلی است. از این رو اطلاعات شکل کلی معمولاً برای کاهش دامنه جستجو در یک سامانه سلسله مراتبی به کار می‌رود [۱۲] تا [۱۴]. کلمات موجود در مجموعه داده بر اساس ویژگی‌های شکل کلی به خوشه‌های مختلف تقسیم شده و در مرحله طبقه‌بندی، هر کلمه ورودی با این خوشه‌ها مقایسه و خوشه متناظر با آن انتخاب می‌شود. بازشناسی نهایی از بین نمونه‌های موجود در این خوشه انجام می‌گردد. کاهش فضای جستجو، علاوه بر کاهش حجم محاسبات مورد نیاز در مراحل بعد، دقت نهایی سامانه بازشناسی را نیز افزایش خواهد داد.

در روش‌های سلسله مراتبی، احتمال وقوع خطا در تعیین خوشه متناظر با کلمه ورودی وجود دارد و با توجه به این که خطای این مرحله به مراحل بعدی انتشار خواهد یافت، طبقه‌بندی باید با دقت بالایی انجام شود. برای افزایش دقت طبقه‌بندی، معمولاً از یک توصیف ساختاری ساده و پایدار برای توصیف شکل کلمات استفاده می‌شود [۱۴]. استفاده از توصیف‌گرهای ساده اگرچه احتمال وقوع خطا را کاهش می‌دهد، اما از آنجا که تفاوت شکل زیر- کلمات را به خوبی نشان نمی‌دهد، نمی‌تواند کاهش چندانی در اندازه دیکشنری ایجاد کند. از سوی دیگر با استفاده از توصیف‌گرهای پیچیده‌تر، احتمال خطا در تشخیص خوشه متناظر با نمونه ورودی نیز بالاتر می‌رود. برای کاهش این خطا، معمولاً به جای یک خوشه، اعضای چند خوشه نزدیک‌تر برای بررسی در مراحل بعدی انتخاب می‌شوند [۱۵].

در این مقاله روشی را برای کاهش اندازه دیکشنری زیر- کلمات چاپی فارسی ارائه می‌کنیم که با حفظ دقت طبقه‌بندی، فضای جستجو را تا حد قابل توجهی کاهش می‌دهد. شکل ۱ اجزای اصلی سامانه پیشنهادی را نشان می‌دهد. ابتدا مجموعه زیر- کلمات پایگاه داده بر اساس ویژگی‌های سراسری شکل به ۳۰۰ خوشه طبقه‌بندی شده و هر کلمه ورودی به مجموعه این خوشه‌های اولیه طبقه‌بندی می‌شود. نتایج به دست آمده از مرحله طبقه‌بندی در مرحله تأیید بررسی می‌گردند. در این مرحله، میزان اطمینان به زیر-کلمات بر اساس ویژگی‌های محلی شکل سنجیده می‌شود و در نهایت با اعمال مجموعه‌ای از قوانین، زیر- کلمات نهایی انتخاب شده و اندازه دیکشنری کاهش داده می‌شود.

ایده اصلی مقاله بر تعریف مناطق شاخص هر خوشه و به کارگیری آن در مرحله تأیید استوار است. مناطق شاخص هر خوشه، مناطقی از شکل زیر- کلمات هستند که نمونه‌های آن خوشه را از سایر نمونه‌ها متمایز می‌کنند. این مناطق در مرحله آموزش برای هر خوشه محاسبه می‌شود. به طور کلی نوآوری‌های این مقاله را می‌توان به صورت زیر خلاصه کرد:



شکل ۱: اجزای اصلی سامانه پیشنهادی برای کاهش فضای جستجو: طبقه‌بندی، تأیید و تصمیم‌گیری.

- ارائه الگوریتمی برای کاهش اندازه دیکشنری زیر- کلمات فارسی.
- توصیف دقیق شکل زیر- کلمات بر اساس ترکیب روش‌های توصیف سراسری و محلی.
- ارائه روشی برای تعیین "مناطق شاخص زیر- کلمه" و به کارگیری آن در جهت توصیف دقیق‌تر شکل زیر- کلمات.
- بسط مفهوم "مناطق شاخص زیر- کلمات" به "مناطق شاخص خوشه‌ها" و استفاده از آن در مرحله تأیید نتایج طبقه‌بندی اولیه.

در این مقاله تنها به بررسی شکل بدنه زیر- کلمات خواهیم پرداخت و نقاط را در نظر نمی‌گیریم. با حذف نقاط از مجموعه ۱۲۷۰۰ زیر- کلمه رایج در زبان فارسی [۶]، تعداد بدنه زیر- کلمات به ۶۸۹۵ کاهش می‌یابد. مجموعه داده‌ای که در این مقاله استفاده می‌شود، تصاویر بدنه ۶۸۹۵ زیر- کلمه است که با قلم لوتوس ۱۴ نگارش و چاپ شده و با درجه تفکیک ۴۰۰ نقطه در اینچ روبش شده‌اند (بخشی از مجموعه داده [۶]).

در بخش ۲ به مرور روش‌های موجود می‌پردازیم و در بخش ۳ ساختار کلی سامانه پیشنهادی ارائه می‌شود. خوشه‌بندی مجموعه‌ی زیر- کلمات و طبقه‌بندی زیر- کلمه ورودی بر اساس توصیف سراسری شکل آن در بخش ۴ شرح داده می‌شود. با این طبقه‌بندی، فضای جستجو برای زیر- کلمه ورودی به اعضای چند خوشه نزدیک‌تر به آن کاهش داده شده و این خوشه‌های انتخابی در سامانه تأیید بررسی می‌شوند. روش تأیید در بخش ۵ آمده است. هدف از بررسی خوشه‌ها در این مرحله، کاهش فضای جستجو به خوشه‌های مطمئن‌تر است. در بخش ۶ چگونگی کاهش اندازه دیکشنری بر اساس نتایج مرحله تأیید بیان می‌شود. ارزیابی روش پیشنهادی و نتایج تجربی در بخش ۷ آمده و در نهایت در بخش ۸ به جمع‌بندی و نتیجه‌گیری می‌پردازیم.

۲- روش‌های موجود

مقالاتی که به بررسی روش‌های توصیف شکل کلی کلمات در زبان‌های مختلف پرداخته‌اند، برای استخراج ویژگی از روش‌های عمومی توصیف شکل استفاده می‌کنند. استخراج ویژگی‌های سراسری از شکل کلمات با استفاده از انواع توصیف‌گرهای ساختاری و آماری مانند افکنش‌های افقی و عمودی، تعداد و موقعیت نقاط و علایم، بالارونده‌ها، پایین‌رونده‌ها، حفره‌ها، گودی‌ها، پروفیل‌های بالا و پایین شکل، تراکم پیکسل‌ها و تبدیل‌های سراسری در مقالات بررسی شده است [۱۶].

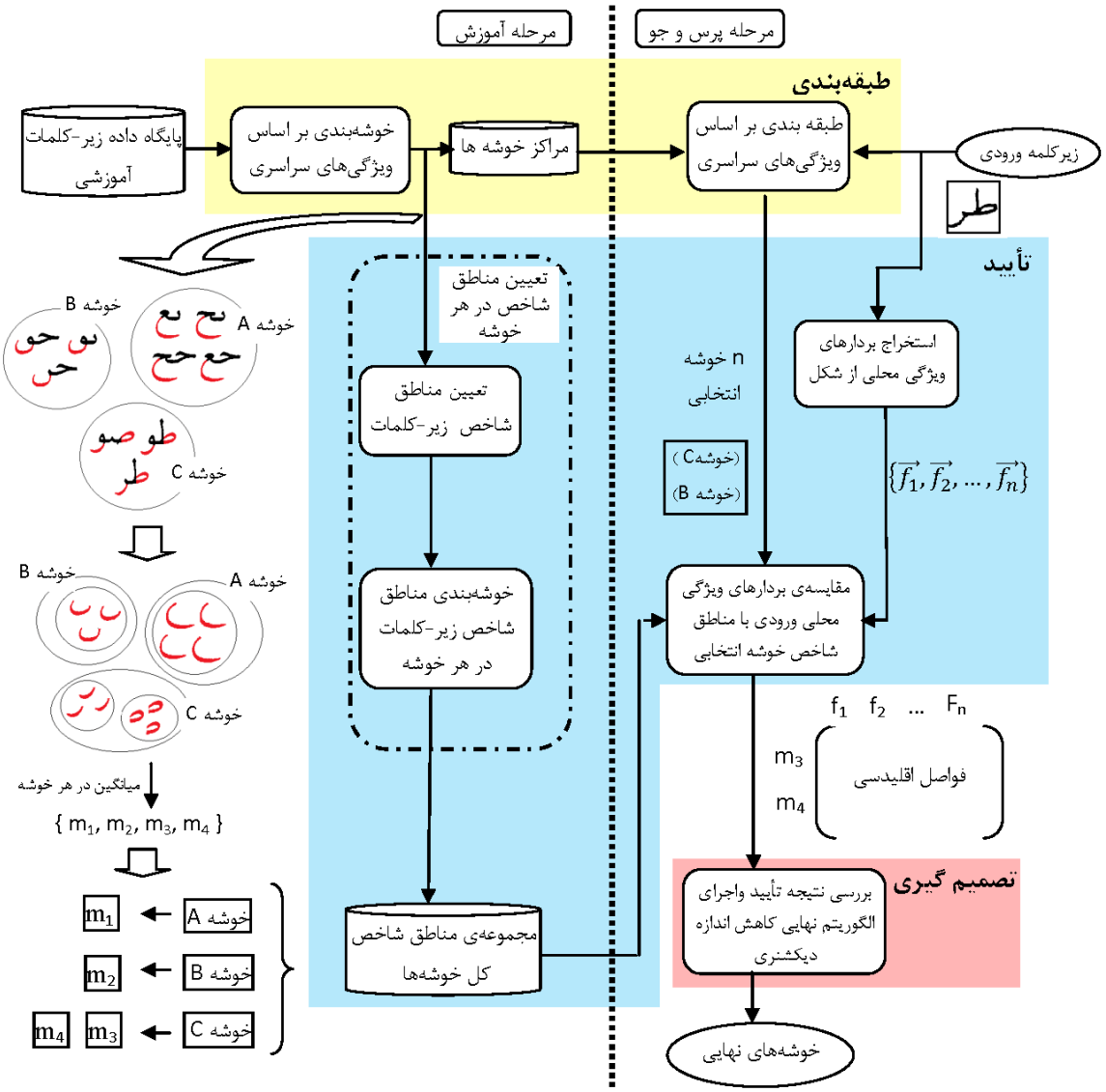
پیوستگی حروف در برخی خطها و در بعضی از شیوه‌های نگارش، قطعه‌بندی شکل کلمه را پیچیده‌تر می‌سازد و موجب می‌شود که محققان گرایش بیشتری به سمت روش‌های مبتنی بر شکل کلی داشته باشند. خطوط فارسی، عربی و همچنین دست‌نویس انگلیسی از این دسته هستند. تحقیقات متعددی درباره بازشناسی خطوط پیوسته انجام شده است [۱۷]. در [۱۶] روش‌های مختلف ارائه‌شده برای بازشناسی کلمات دست‌نویس با رویکرد مبتنی بر شکل کلی بررسی شده و [۱۸] به مرور روش‌های بازشناسی کلمات عربی با هر دو رویکرد مبتنی بر قطعه‌بندی و مبتنی بر شکل کلی پرداخته است.

با توجه به پیوسته‌نویسی در خط فارسی، ویژگی‌های مبتنی بر شکل کلی برای توصیف کلمات فارسی مناسب می‌باشد، با این حال تحقیقات محدودی در این زمینه انجام گرفته است. نتایج همین تحقیقات، حکایت از کارامدی این روش‌ها در توصیف کلمات فارسی دارد [۱۲]، [۱۵] و [۱۹] تا [۲۲]. در [۲۲] از ۴۵ گشتاور زرنیکی برای توصیف زیر- کلمات چاپی و دست‌نویس فارسی استفاده شده است. در [۲۱] از مدل پنهان مارکوف برای مدل‌کردن شکل کلمات استفاده شده است. برای تعیین ویژگی‌ها در هر قاب تصویر، کد زنجیره‌ای کانتورشکل در خانه‌های قاب محاسبه می‌شود. مرجع [۲۰] برای توصیف شکل کلمات به بررسی موقعیت و طول پاره مسیرهای روی کانتور بالایی نسبت به خط زمینه پرداخته و برای هر نقطه روی کانتور برچسبی در نظر می‌گیرد. برای تعیین همسایگی‌های هر کلمه ورودی، آن کلمه بر اساس برچسب‌های کانتور با کلمات دیکشنری مقایسه می‌شود. در [۱۵] روشی دومرحله‌ای برای بازشناسی زیر- کلمات چاپی فارسی ارائه شده است. در مرحله اول با استفاده از ویژگی‌های مکان مشخصه، مجموعه زیر- کلمات به تعدادی خوشه تقسیم می‌شوند و در مرحله دوم، هر کلمه با استفاده از توصیف‌گرهای فوریه کانتور آن توصیف می‌شود.

۳- ساختار کلی روش پیشنهادی

شکل ۲ ساختار روش پیشنهادی را نشان می‌دهد و همان طور که در بخش قبل اشاره شد، این ساختار از سه بخش اصلی طبقه‌بندی، تأیید و تصمیم‌گیری تشکیل شده است. در مرحله طبقه‌بندی از ویژگی‌های سراسری شکل برای توصیف زیر- کلمات استفاده می‌شود. هر زیر- کلمه ورودی ابتدا با مراکز خوشه‌هایی که در مرحله آموزش به دست آمده‌اند، مقایسه شده و فهرستی از n خوشه نزدیک‌تر انتخاب می‌شود. خوشه‌ها با روش k- میانگین ایجاد شده‌اند و طبقه‌بندی بر اساس معیار فاصله اقلیدسی از مراکز خوشه‌ها انجام می‌شود.

سپس در مرحله تأیید از ویژگی‌های محلی شکل زیر- کلمه برای بررسی نتیجه طبقه‌بندی استفاده می‌شود. مناطق محلی در تصویر زیر- کلمه ورودی با مناطق شاخص خوشه مقایسه می‌شود. مناطق شاخص برای تمام خوشه‌ها پیش از این و در مرحله آموزش تعیین و ذخیره شده و هر خوشه بر اساس ویژگی‌های محلی اعضای آن نسبت به اعضای سایر خوشه‌ها با تعدادی مناطق شاخص بازنمایی شده است. مناطق شاخص هر خوشه، بخش‌هایی از شکل اعضای آن هستند که قابلیت بیشتری در ایجاد تمایز بین نمونه‌های آن خوشه و سایر نمونه‌ها دارند. تعیین مناطق شاخص هر خوشه در مرحله آموزش از دو قسمت تشکیل شده است: ۱) تعیین مناطق شاخص زیر- کلمات و ۲) گروه‌بندی مناطق شاخص زیر- کلمات هر خوشه.



شکل ۲: فرایند کلی سامانه طبقه‌بندی زیر- کلمات بر اساس بخش‌های متمایزکننده هر خوشه.

روشی که برای تعیین مناطق شاخص زیر- کلمات به کار می‌گیریم از روش ارائه‌شده در [۲۳] الهام گرفته شده که در اصل برای شکل‌های سه‌بعدی ارائه شده و در بخش ۵-۲-۱ شرح داده می‌شود. بعد از تعیین مناطق شاخص در شکل زیر- کلمات، از آنها برای استخراج "مناطق شاخص خوشه" استفاده می‌شود و برای این منظور، مناطق شاخص زیر- کلمات عضو یک خوشه گروه‌بندی می‌شوند. میانگین اعضای گروه‌های ایجادشده، مناطق شاخص آن خوشه را تشکیل می‌دهند.

در شکل ۲ برای تشریح روش تأیید، مثالی از تصاویر بدنه ۱۰ زیر- کلمه دوحرفی آمده است که در ۳ خوشه قرار گرفته‌اند و در بخش ۵-۲ از این مثال برای بیان روشن‌تر مفهوم مناطق شاخص نیز استفاده خواهیم کرد. مناطق شاخص زیر- کلمات در این مثال با رنگ قرمز مشخص شده‌اند و با خوشه‌بندی این مناطق در هر خوشه، مناطق شاخص برای هر سه خوشه تعیین شده است.

بر اساس نتایج مقایسه مناطق محلی زیر- کلمه ورودی با مناطق شاخص خوشه، درباره تأیید یا رد آن خوشه تصمیم‌گیری می‌شود. اگر شرط شباهت کافی بین مناطق شاخص خوشه و مناطق محلی زیر- کلمه‌ی ورودی برآورده شود آن خوشه، خوشه مطمئنی است و به عنوان خوشه نهایی پذیرفته شده و در غیر این صورت، خوشه بعدی در مرحله تأیید بررسی می‌شود. بررسی خوشه‌های حاصل از طبقه‌بندی تا جایی ادامه می‌یابد که خوشه‌ای در مرحله تأیید به عنوان خوشه مطمئن شناخته شود. به این ترتیب فضای جستجو برای کلمه ورودی به اعضای آن خوشه و خوشه‌های نزدیک‌تر محدود می‌شود. در بخش‌های بعد، اجزای مختلف ساختار پیشنهادی ارائه می‌شود.

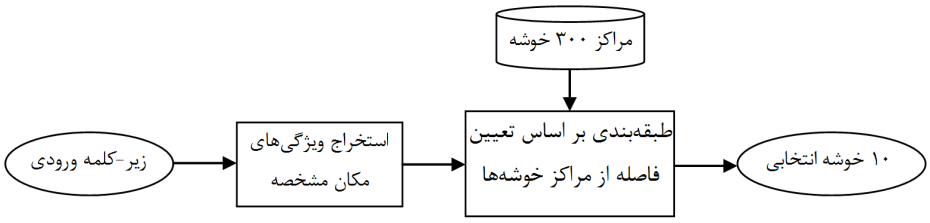
٤- طبقه‌بندی بر اساس ویژگی‌های سراسری

ابتدا زیر- کلمات موجود در پایگاه داده در مرحله طبقه‌بندی خوشه‌بندی می‌شوند. نمونه ورودی به این خوشه‌ها طبقه‌بندی شده و فهرستی از خوشه‌های نزدیک‌تر به آن انتخاب می‌شود. نتیجه این طبقه‌بندی در مراحل بعد بررسی می‌شود و خوشه‌های نهایی انتخاب می‌شوند.

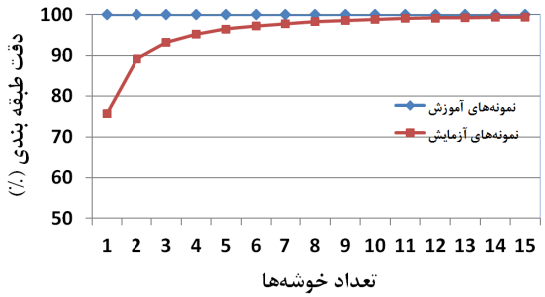
برای طبقه‌بندی بر اساس ویژگی‌های سراسری، روش ارائه‌شده در [۴] را به کار می‌بریم. شکل ۳ بلوک دیاگرام طبقه‌بندی بر اساس ویژگی‌های سراسری را نشان می‌دهد.

٤-١ استخراج ویژگی‌های مکان مشخصه

برای توصیف شکل زیر- کلمات از ویژگی‌های مکان مشخصه استفاده می‌کنیم [۴] و [۱۵]. این توصیف‌گر به هر نقطه پس‌زمینه تصویر، یک کد نسبت داده و فراوانی نقاط با کدهای یکسان، بردار ویژگی نهایی را تشکیل می‌دهد.



شکل ۳: بلوک دیاگرام طبقه‌بندی زیر- کلمه ورودی بر اساس ویژگی‌های سراسری به ۳۰۰ خوشه.



شکل ۴: نتایج طبقه‌بندی زیر- کلمات بر اساس ویژگی‌های مکان مشخصه به ۳۰۰ خوشه در ۱۵ انتخاب اول.

برای تعیین کد متناظر با هر نقطه، از آن نقطه دو خط عمودی و افقی رسم شده و تعداد برخورد این خطوط با بدنه تصویر در چهار جهت راست، بالا، چپ و پایین تعیین می‌شود. بیشینه تعداد برخوردها را به ۳ محدود می‌کنیم. از کنار هم قرار دادن ۴ عدد به دست آمده، یک عدد ۴ رقمی در مبنای ۴ به دست خواهد آمد که مقدار این عدد در مبنای ۱۰، کد متناظر با آن نقطه پس‌زمینه خواهد بود. به این ترتیب، بردار ویژگی‌های مکان مشخصه هر شکل، برداری ۲۵۶بعدی است که هر بعد آن، فراوانی نقاط با کد متناظر با آن بعد را نشان می‌دهد. برای نرمال‌کردن این ویژگی‌ها، مقادیر به دست آمده به تعداد کل نقاط پس‌زمینه تقسیم و سپس بردارهای ویژگی با روش تحلیل مؤلفه‌های اصلی به ۲۵ کاهش داده می‌شود.

٤-۲ خوشه‌بندی مجموعه زیر- کلمات و طبقه‌بندی زیر- کلمه ورودی

پس از استخراج ویژگی‌های مکان مشخصه از تصاویر زیر- کلمات، نمونه‌ها با استفاده از الگوریتم k میانگین خوشه‌بندی می‌شوند. برای خوشه‌بندی از معیار فاصله اقلیدسی استفاده می‌شود و تصاویر ۶۸۹۵ زیر- کلمه به ۳۰۰ خوشه تقسیم می‌شوند. با این تقسیم‌بندی، کمینه و بیشینه تعداد اعضای هر خوشه ۱ و ۵۴ نمونه است. میانگین بردارهای ویژگی در هر خوشه به عنوان نماینده آن خوشه در نظر گرفته می‌شود.

برای طبقه‌بندی زیر- کلمه ورودی، ابتدا ویژگی‌های مکان مشخصه از تصویر آن استخراج و این بردار با نماینده‌های خوشه‌ها مقایسه می‌شود و خوشه‌های نزدیک‌تر مشخص می‌گردند. شکل ۴ نمودار دقت طبقه‌بندی زیر- کلمات را در ۱۵ انتخاب اول نشان می‌دهد و در این شکل دو نمودار به ازای نمونه‌های آموزشی و آزمایشی رسم شده است. ۶۸۹۵ نمونه‌ای که برای ایجاد خوشه‌ها به کار رفته‌اند به عنوان نمونه‌های آموزشی استفاده شده‌اند. برای ایجاد نمونه‌های آزمایشی نیز از مجموع تصاویر ۵۰۰۰ زیر- کلمه استفاده شده که در اندازه‌های مختلف چاپ و با درجات تفکیک متفاوت روبش شده است. با توجه به این شکل با انتخاب ۱۰ خوشه نزدیک‌تر، برای مجموعه آزمایش به دقتی نزدیک به ۱۰۰٪ دست خواهیم یافت، از این رو تعداد خوشه‌های انتخابی در مرحله طبقه‌بندی را ۱۰ خوشه اول در نظر می‌گیریم و به این ترتیب، زیر- کلمه متناظر با ورودی به احتمال زیاد در بین اعضای این ۱۰ خوشه موجود است.

۵- تأیید بر اساس مقایسه ویژگی‌های محلی زیر- کلمه با مناطق شاخص خوشه

بعد از طبقه‌بندی زیر- کلمه ورودی، خوشه‌های به دست آمده در فرایند تأیید بررسی می‌شوند و در این بخش جزئیات الگوریتم تأیید پیشنهادی را شرح می‌دهیم. در روش پیشنهادی از ویژگی‌های محلی شکل برای سنجش میزان اطمینان به خوشه انتخابی استفاده می‌شود. بردارهای ویژگی از مناطق محلی شکل ورودی استخراج و با مناطق شاخص هر خوشه مقایسه می‌شود. مناطق شاخص هر خوشه، دسته‌ای از مناطق محلی شکل هستند که در بین اعضای آن خوشه فراوان‌ترند و به این دلیل به عنوان مشخصه آن خوشه در نظر گرفته می‌شوند. بر این اساس اگر تصویر یک زیر- کلمه شامل مناطق شاخص یک خوشه باشد، می‌توان با احتمال بالایی آن زیر- کلمه را عضو آن خوشه دانست. مناطق شاخص هر خوشه در مرحله آموزش تعیین شده و به آن نسبت داده می‌شود.

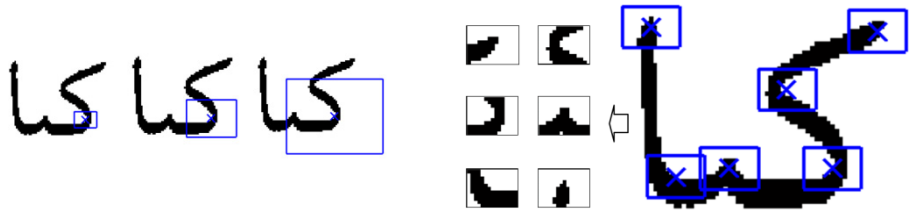
۵-۱ منطقه‌بندی تصویر و استخراج ویژگی از هر منطقه

برای منطقه‌بندی شکل هر زیر- کلمه، تعدادی نقاط ویژه از شکل استخراج می‌شود. پنجره‌ای با اندازه معین به دور هر نقطه قرار داده شده و یک منطقه از شکل مشخص می‌شود و با استخراج ویژگی از هر منطقه، بردار ویژگی آن منطقه محاسبه می‌شود.

برای تعیین نقاط ویژه در تصویر هر زیر- کلمه، از آشکارساز گوشه هریس [۲۴] استفاده کرده‌ایم. گوشه‌ها در شکل زیر- کلمات فارسی از ویژگی‌های حائز اهمیت شکل هستند و تکرارپذیری قابل قبولی از خود نشان می‌دهند.

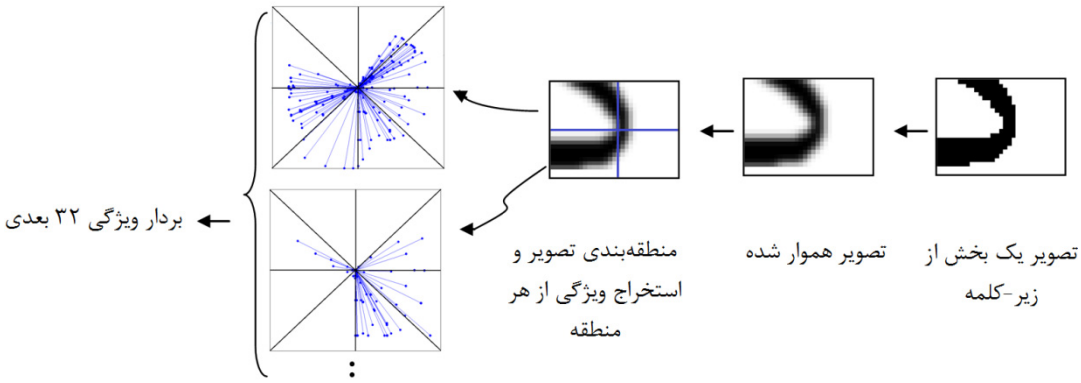
یک قاب مستطیل‌شکل به مرکز هر کدام از نقاط ویژه روی تصویر قرار داده می‌شود. نواحی ایجادشده به این طریق، مناطق محلی شکل زیر- کلمه را می‌سازند و اندازه این قاب متناسب با اندازه مستطیل محیطی شکل زیر- کلمه انتخاب می‌شود. سه حالت مختلف برای اندازه قاب در هر شکل در نظر می‌گیریم. اندازه طول و عرض هر قاب در این سه حالت به ترتیب ۱/۴، ۱/۲ و ۱ برابر اندازه طول و عرض مستطیل محیطی تصویر زیر- کلمه در نظر گرفته شده و شکل ۵ مناطق ایجادشده در شکل یک زیر- کلمه را نشان می‌دهد. در شکل ۵- الف تمام بخش‌ها با مقیاس ۱/۴ از شکل کلمه استخراج شده و در شکل ۵- ب سه قاب با اندازه‌های ۱/۴، ۱/۲ و ۱ برای استخراج سه بخش از یک نقطه ویژه به کار گرفته شده است. در ادامه به جز در شرایطی که اندازه مقیاس ذکر شده باشد، مقیاس پیش‌فرض برای استخراج نواحی ۱/۲ در نظر گرفته شده است. در شرایطی که بخشی از قاب خارج از محدوده تصویر زیر- کلمه قرار بگیرد، اندازه‌های پیکسل‌ها در آن بخش صفر در نظر گرفته می‌شود.

پس از استخراج مناطق محلی از تصویر یک زیر- کلمه، هر کدام از این مناطق با یک بردار ویژگی توصیف می‌شود. برای توصیف هر منطقه، توصیف‌گر هیستوگرام جهت گرادیان انتخاب شده [۲۵] و این توصیف‌گر، چگونگی توزیع اندازه و جهت گرادیان نقاط تصویر را در نواحی مختلف

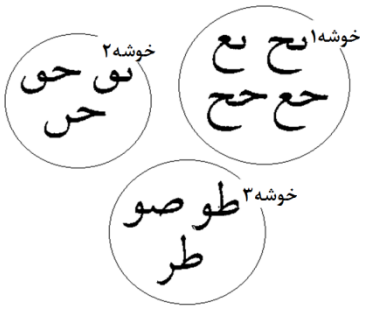


(الف) (ب)

شکل ۵: (الف) استخراج مناطق محلی از شکل یک زیر–کلمه نمونه (مقیاس پنجره ۱/۴ در نظر گرفته شده) و (ب) سه منطقه محلی در یک نقطه ویژه با سه مقیاس ۱/۴، ۱/۲ و ۱.



شکل ۶: فرایند استخراج ویژگی هیستوگرام جهات گرادیان از یک منطقه محلی.



شکل ۷: خوشه‌بندی بدنه ۱۰ زیر– کلمه نمونه به ۳ خوشه.

شکل نشان می‌دهد. برای محاسبه این ویژگی (شکل ۶) تصویر پس از هموارسازی به ۴ ناحیه تقسیم می‌شود. اندازه و جهت گرادیان در هر نقطه محاسبه شده و هیستوگرام جهت‌های گرادیان در هر کدام از نواحی ۴گانه تصویر محاسبه می‌شود. ۸ جهت را برای محاسبه هیستوگرام در هر ناحیه در نظر گرفته‌ایم که این ۸ جهت از ۰ تا ۳۱۵ درجه و به فاصله ۴۵ درجه از هم تعیین شده‌اند. برای هر پیکسل در هر ناحیه، دو زاویه نزدیک‌تر به زاویه گرادیان تعیین می‌شود. اندازه گرادیان در آن پیکسل، متناسب با فاصله‌اش از این دو جهت، بین آنها تقسیم می‌شود و هیستوگرام اندازه‌ها در زوایای مختلف محاسبه می‌شود. با کنار هم قرار دادن هیستوگرام‌های محاسبه‌شده در ۴ ناحیه، بردار ویژگی ایجاد می‌شود و به این ترتیب برای هر منطقه از شکل زیر–کلمه، توصیف‌گری به طول $۸\times۴=۳۲$ به دست می‌آید. بردار توصیف‌گر هر منطقه به اندازه یک نرمال شده است.

۵-۲ تعیین مناطق شاخص هر خوشه

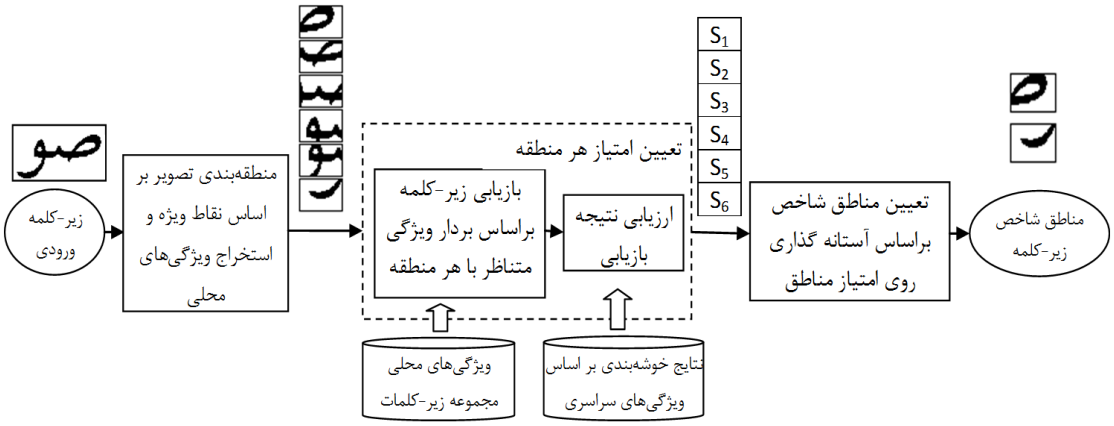
منظور از مناطق شاخص هر خوشه، بخش‌هایی از شکل است که بین نمونه‌های متعلق به آن خوشه بیشترین شباهت را دارند و در عین حال بیشترین تفاوت را با نمونه‌های سایر خوشه‌ها ایجاد می‌کنند. برای بیان روشن‌تر به شکل ۷ توجه کنید.

در این شکل، تصاویر بدنه ۱۰ زیر– کلمه دو حرفی در ۳ خوشه قرار گرفته است. انتخاب این کلمات و نحوه قرارگرفتن آنها در خوشه‌ها، تنها برای بیان بهتر مفهوم "بخش‌های متمایزکننده هر خوشه" است. در خوشه ۱ همه زیر–کلمات به حروف "ع" یا "ح" ختم می‌شوند و این حروف در زیر–کلمات سایر خوشه‌ها دیده نمی‌شود. دایره این حروف می‌تواند به عنوان منطقه‌ای شاخص در این خوشه در نظر گرفته شود. برای ایجاد تمایز بین زیر– کلمات این خوشه و خوشه ۳ از شکل حروف اول نیز می‌توان استفاده کرد اما این حروف با حروف اول نمونه‌های موجود در خوشه ۲ مشابه است و از این رو از نظر اهمیت در رده بعدی قرار می‌گیرند. در خوشه ۳ بخش منحنی شکل حروف "ر" و "و" می‌تواند به عنوان منطقه‌ای شاخص در این خوشه انتخاب شود. دقت کنید که حفره موجود در حرف "و" به دلیل شباهت با حفره حرف "ق" برای ایجاد تمایز بین نمونه‌های خوشه ۳ و ۲ مناسب نیست. حفره موجود در حروف "ط" و "صـ" نیز نمونه‌های خوشه ۳ را از سایر نمونه‌ها متمایز می‌کند. دسته "ط" در دو زیر– کلمه این خوشه وجود دارد و این دو زیر–کلمه را از نمونه‌های سایر خوشه‌ها جدا می‌کند. بر این اساس، این بخش هم برای این گروه شاخص است اما از نظر اهمیت در مرتبه‌های بعد قرار می‌گیرد.

۵-۲-۱ تعیین مناطق شاخص هر خوشه

برای تعیین مناطق شاخص خوشه‌ها، ابتدا مناطق شاخص در هر زیر–کلمه تعیین می‌شود. مناطق شاخص هر زیر– کلمه، بخش‌هایی از شکل آن زیر– کلمه است که در زیر– کلمات هم‌خوشه فراوان دیده می‌شود، در حالی که احتمال یافتن آن بخش در سایر زیر– کلمات کم است. پس از استخراج مناطق محلی از شکل زیر–کلمات، بنا بر تعریف ارائه‌شده، تعدادی از این مناطق به عنوان مناطق شاخص انتخاب می‌شوند. روش انتخاب مناطق شاخص، بر نتیجه بازیابی شکل زیر–کلمه بر اساس اطلاعات آن بخش استوار است. به این منظور، مناطق مختلف شکل با معیارهای مبتنی بر نتیجه بازیابی، ارزیابی شده و مناطق شاخص تعیین می‌شوند.

ایده تعیین مناطق شاخص زیر– کلمات از [۲۳] الهام گرفته شده که در اصل برای شکل‌های سه‌بعدی ارائه شده است. شکل ۸ روندنمای تعیین



شکل ۸: روندنمای تعیین مناطق شاخص یک زیر- کلمه.

مناطق شاخص در شکل یک زیر- کلمه را نشان می‌دهد. پس از قطعه‌بندی شکل زیر- کلمه و استخراج بردارهای ویژگی محلی، به هر منطقه محلی امتیازی نسبت داده می‌شود که نشان می‌دهد آن منطقه از شکل تا چه اندازه به زیر- کلمات هم‌خوشه منحصر است. برای بررسی این موضوع از اطلاعات شکل هر منطقه برای بازیابی زیر- کلمه متناظر در پایگاه داده استفاده می‌شود. نتیجه مطلوب این بازیابی، نتیجه‌ای است که همه زیر- کلمات هم‌خوشه با آن زیر- کلمه را در اول فهرست بازیابی قرار دهد. در این صورت می‌توان نتیجه گرفت که زیر- کلمات هم‌خوشه، از جهت دارابودن این منطقه از شکل، بیشترین شباهت را داشته و با زیر- کلمات سایر خوشه‌ها متفاوتند. به این ترتیب با ارزیابی نتیجه بازیابی، به هر منطقه از زیر- کلمه امتیازی اختصاص داده می‌شود و پرامتیازترین مناطق به عنوان مناطق شاخص زیر- کلمه انتخاب می‌شوند.

برای بررسی مناطق محلی در هر زیر- کلمه، روشی را که در [۲۶] معرفی کرده‌ایم با اندکی تغییرات به کار می‌گیریم. تفاوت عمده روش این مقاله با [۲۶] در تعریف کلاس منتسب به هر زیر- کلمه است. در [۲۶] هر زیر- کلمه به همراه نمونه‌های مختلف نگارش آن، یک کلاس را تشکیل می‌دهند اما در اینجا کلاس‌ها، همان خوشه‌هایی هستند که از توصیف سراسری شکل کلمات به دست می‌آیند. بنابراین اعضای یک کلاس لزوماً نمونه‌هایی از یک زیر- کلمه مشخص نیستند بلکه زیر- کلماتی هستند که از نظر شکل کلی شباهت‌هایی دارند. در این حالت، مناطق شاخص در شکل هر زیر- کلمه، بخش‌هایی از شکل هستند که به خوشه‌بندی درست آن زیر- کلمه کمک می‌کنند. در ادامه این بخش روش تعیین امتیاز مناطق مختلف زیر- کلمه و تعیین مناطق شاخص آن را شرح می‌دهیم.

$\{M_1, M_2, \ldots, M_n\}$ مجموعه شکل زیر- کلمات موجود در پایگاه داده است و x_{ij}، iامین بردار ویژگی محلی در زیر- کلمه M_j را نشان می‌دهد. برای تعیین امتیاز هر منطقه با بردار x_{ij}، فرض می‌کنیم که این بردار ویژگی تنها اطلاعاتی است که از شکل M_j در اختیار داریم و تنها بر اساس همین اطلاعات به جستجوی زیر- کلمات هم‌خوشه با زیر- کلمه M_j می‌پردازیم. این کار برای مناطق مختلف زیر- کلمه M_j انجام شده و امتیاز هر منطقه x_{ij} از زیر- کلمه M_j به صورت زیر محاسبه می‌شود:

۱) فاصله بین x_{ij} از یک زیر- کلمه با هر زیر- کلمه دیگر M_t با استفاده از (۱) تعیین می‌شود

$$dist(x_{ij}, M_t) = \min_b (d(x_{ij}, x_{bt})) \qquad (۱)$$

که $d(.)$ فاصله اقلیدسی بین دو بردار ویژگی را نشان می‌دهد.

۲) فاصله‌های محاسبه‌شده از کوچک به بزرگ مرتب می‌شوند و فهرست زیر- کلمات متناظر با این فواصل، نتیجه بازیابی زیر- کلمات هم‌خوشه با M_j با استفاده از بردار ویژگی x_{ij} است.

۳) نتیجه بازیابی ارزیابی شده و امتیاز منطقه مربوط تعیین می‌شود. هر چقدر زیر- کلمات هم‌خوشه با M_j در قسمت‌های بالاتر فهرست قرار بگیرند، بدین معنی است که منطقه متناظر با x_{ij} تمایز بیشتری بین نمونه‌های هم‌خوشه با M_j و سایر زیر- کلمات ایجاد می‌کند.

برای ارزیابی نتیجه بازیابی هر زیر- کلمه از معیار DCG[1] استفاده شده است. برای محاسبه معیار DCG ابتدا برداری باینری (G) به فهرست بازیابی‌شده نسبت داده می‌شود. مؤلفه‌هایی از این بردار که متناظر با نمونه‌های هم‌خوشه با پرس و جو هستند، یک و سایر مؤلفه‌ها صفر در نظر گرفته می‌شود. معیار DCG از (۲) محاسبه می‌شود

$$DCG_i = \begin{cases} G_1 & , \quad i = 1 \\ DCG_{i-1} + \dfrac{G_i}{\log_2(i)} & , \quad \text{otherwise} \end{cases} \qquad (۲)$$

نتیجه نهایی به بیشینه مقدار ممکن برای DCG نرمال می‌شود. هر چقدر مقدار DCG محاسبه‌شده به یک نزدیک‌تر باشد، بدین معنی است که بازیابی نتیجه بهتری داشته است. امتیازی که به هر بردار ویژگی و منطقه متناظر با آن در شکل زیر- کلمه اختصاص داده می‌شود، همین مقدار DCG به دست آمده از بازیابی است.

اصلاح روش امتیازدهی

با توجه به تعداد زیاد خوشه‌ها (۳۰۰) و زیر- کلمات (۶۸۹۵)، تعیین امتیاز با روش فوق با محدودیت‌هایی همراه است. وقتی تعداد کل زیر- کلمات افزایش یابد، احتمال حضور نمونه‌های غیر هم‌خوشه در فهرست بازیابی نیز افزایش یافته و باعث کاهش دامنه تغییرات امتیازها می‌شود. در شکل ۹ فهرست بازیابی یک زیر- کلمه نمونه برای یکی از مناطق محلی آن تا چند انتخاب اول نشان داده شده است. همان طور که مشخص است، زیر- کلمات بازیابی شده (که هیچ یک با نمونه ورودی هم‌خوشه نیستند) با وجود شباهت در یک منطقه محلی از نظر ساختار کلی شکل تا حد زیادی با زیر- کلمه پرس و جو متفاوتند. با توجه به این که پیش از مرحله تأیید، هر نمونه ورودی به خوشه‌هایی طبقه‌بندی می‌شود که از نظر ویژگی‌های سراسری مشابه هستند، در اینجا نیز برای تعیین امتیاز مناطق یک زیر- کلمه، فقط به بررسی خوشه‌های مجاور با آن می‌پردازیم.

1. Discounted Cumulative Gain

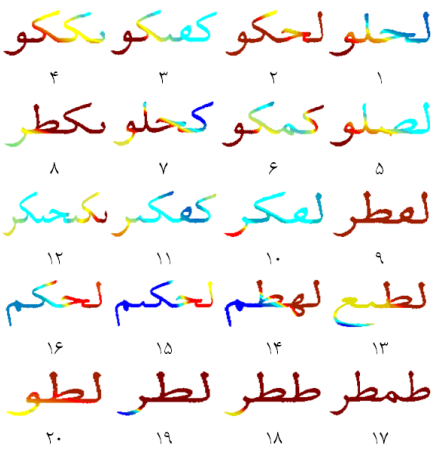
شکل ۹: نتیجه بازیابی زیر– کلمه پرس و جو بر اساس یک منطقه محلی (مناطق متناظر در هر شکل با مستطیل نشان داده شده است).

شکل ۱۰: نتیجه بازیابی زیر– کلمه پرس و جو بر اساس یک منطقه محلی. بازیابی بین اعضای ۲۰ خوشه نزدیک‌تر انجام شده و زیر– کلماتی که با رنگ آبی انتخاب شده‌اند با پرس و جو هم‌خوشه هستند.

بر این اساس برای تعیین امتیاز هر منطقه از زیر–کلمه، ابتدا زیر–کلمه ورودی از بین اعضای این خوشه‌ها بازیابی می‌شود و مقدار n به طور تجربی ۲۰ انتخاب شده است. مقادیر امتیازها در روش اصلاح‌شده در محدوده وسیع‌تری توزیع می‌شود و شکل ۱۰ فهرست بازیابی را برای همان مثال شکل ۹ با روش اصلاحی نشان می‌دهد.

الگوریتم فوق را برای تمام زیر– کلمات اجرا کرده‌ایم و امتیاز مناطق مختلف زیر– کلمات محاسبه شده است. شکل ۱۱ تعدادی از زیر–کلمات عضو یک خوشه را همراه با امتیاز مناطق محلی آنها نشان می‌دهد و امتیاز مناطق محلی با رنگ آنها مشخص شده است. طیف رنگ از آبی به سمت قرمز افزایش امتیاز را نشان می‌دهد. شکل ۱۱ تعدادی از زیر–کلمات عضو یک خوشه را همراه با امتیاز مناطق محلی آنها نشان می‌دهد و امتیاز مناطق محلی با رنگ آنها مشخص شده است. طیف رنگ از آبی به سمت قرمز افزایش امتیاز را نشان می‌دهد و به این ترتیب، بخش‌های قرمزرنگ پرامتیازترین و بخش‌های آبی‌رنگ کم‌امتیازترین بخش‌ها هستند. در ترسیم این شکل برای این که بتوانیم امتیاز بخش‌های هم‌پوشان را بهتر نمایش دهیم به جای هر منطقه، به هر نقطه از شکل امتیازی نسبت داده شده است که امتیاز هر نقطه، رنگ متناظر با آن را مشخص می‌کند و امتیاز نقاط از ترکیب وزن‌دار امتیازهای محاسبه‌شده در نقاط ویژه به دست می‌آید. وزن هر نقطه بر اساس فاصله آن نقطه از نقاط ویژه، تعیین می‌شود. این تخصیص امتیاز به نقاط، تنها به منظور ارائه یک نمایش گویاتر است و در روش پیشنهادی تنها امتیاز مناطق محلی را در نظر خواهیم گرفت.

برای تفسیر امتیاز مناطق مختلف زیر– کلمات در این شکل، نیازمند آن هستیم که از اعضای سایر خوشه‌ها نیز مطلع باشیم. با این حال با بررسی همین شکل به تنهایی هم می‌توان به نکات قابل توجهی دست یافت. بخش‌های انتهایی زیر– کلمات ۱ تا ۷ و همچنین زیر– کلمات ۹ تا ۱۲ که به حروف "و" و "ر" ختم شده‌اند، همگی از مناطق پرامتیاز به حساب می‌آیند. رنگ این حروف در زیر–کلمات ۲، ۳، ۴ و ۶ نزدیک به قرمز است. این به دلیل وجود حرف "ک" پیش از حرف "و" انتهایی است که این دو نمونه را به خوبی از نمونه‌های سایر خوشه‌ها جدا می‌سازد. ترکیب "طر" در انتهای شکل زیر– کلمات ۹، ۱۷، ۱۸ و ۱۹ نیز شرایط مشابهی دارد. امتیاز بیشتر این منطقه در زیر–کلمات ۹ و ۱۷ می‌تواند به دلیل وجود حفره‌های مربوط به حروف پیشین آنها ("فـ" و "مـ") باشد. ترکیب حروف "لط" در ابتدای کلمات ۱۳، ۱۹ و ۲۰ بالاترین امتیاز را در این زیر– کلمات دارد. علاوه بر این سه زیر– کلمه، حرف "لـ" در ابتدای زیر–کلمات ۱۳ تا ۱۶ نیز با بخشی از حروف بعدی به عنوان مناطق شاخص



شکل ۱۱: تعدادی از زیر– کلمات عضو یک خوشه. امتیاز هر منطقه با رنگ آن مشخص شده و طیف رنگ از آبی به سمت قرمز افزایش امتیاز را نشان می‌دهد.

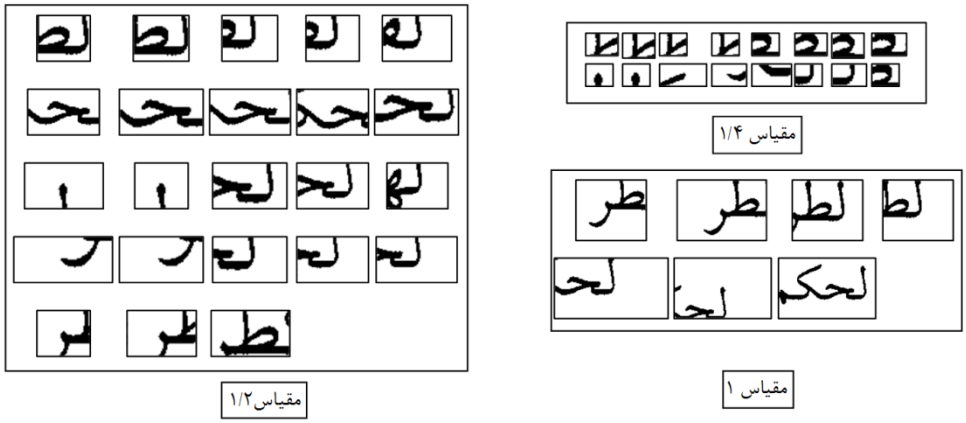
آنها در نظر گرفته شده است. وجود حرف "حـ" و ترکیب آن با بخشی از حروف بعدی در زیر– کلمات ۱، ۲، ۷، ۱۵ و ۱۶ نیز در تشکیل مناطق شاخص این زیر– کلمات مؤثر است. از سویی دیگر، تعدد حروفی چون "ک" یا "م" انتهایی، در زیر– کلمات این گروه (زیر– کلمات ۱۰، ۱۱، ۱۲، ۱۴، ۱۵ و ۱۶) باعث نشده که این بخش‌ها به عنوان مناطق شاخص زیر– کلماتشان در نظر گرفته شوند و این موضوع می‌تواند به دلیل وجود بخش‌های مشابه در زیر– کلمات سایر خوشه‌ها باشد.

پس از تعیین امتیاز مناطق محلی، مناطق شاخص زیر– کلمه تعیین می‌شود و برای تعیین مناطق شاخص زیر– کلمه، آستانه‌ای برای امتیاز مناطق محلی در نظر گرفته می‌شود. مناطقی که امتیازی بیشتر از این آستانه دارند به عنوان مناطق شاخص زیر– کلمه انتخاب می‌شوند. با توجه به این که برای استخراج مناطق محلی، سه مقیاس مختلف در نظر گرفته شده، برای هر مقیاس آستانه جداگانه‌ای تعیین می‌شود. مقدار آستانه‌ها برای مقیاس‌های ۱/۴، ۱/۲ و ۱ به ترتیب ۰/۶، ۰/۷ و ۰/۸ در نظر گرفته شده است.

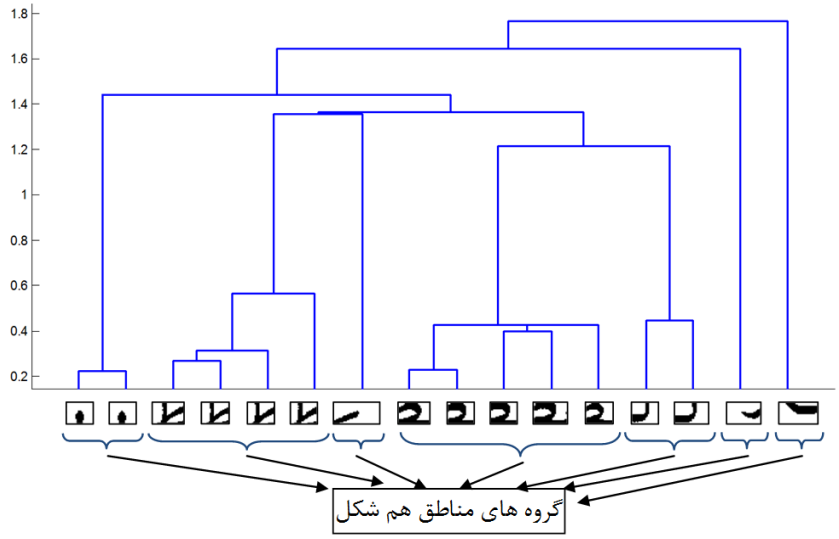
۵-۲-۲ گروه‌بندی مناطق شاخص زیر– کلمات هر خوشه

مناطق شاخص زیر– کلمات یک خوشه در کنار هم قرار گرفته و گروه‌هایی از مناطق هم‌شکل را ایجاد می‌کنند. مناطق شاخص زیر– کلمات عضو یک خوشه در شکل ۱۲ نشان داده شده و این خوشه، همان خوشه‌ایست که در شکل ۱۱ آمده است. در این شکل، مناطق شاخص در هر ۳ مقیاس نشان داده شده و تعداد این مناطق برای خوشه‌های مختلف از ۰ تا ۴۷۳ متغیر است که مقدار صفر بیانگر وجود خوشه‌هایی است که هیچ بخش مهمی ندارند. البته ممکن است که یک خوشه در یک مقیاس خاص بخش مهمی نداشته باشد اما بخش‌های مهمی در مقیاس‌های دیگر داشته باشد. تعداد خوشه‌هایی که هیچ بخش مهمی با مقیاس ۱/۴ ندارند، ۴۱ است که این مقدار برای مقیاس‌های ۱/۲ و ۱ به ترتیب ۸۵ و ۱۷۰ است. میانگین تعداد بخش‌های مهم در خوشه‌ها برای اندازه‌های ۱/۴، ۱/۲ و ۱ به ترتیب ۳۶/۳، ۱۶ و ۹/۸ است.

مناطق شاخص زیر– کلمات در هر خوشه برای ایجاد مناطق شاخص آن خوشه، گروه‌بندی می‌شوند و نماینده گروه‌ها، مناطق شاخص خوشه را تشکیل می‌دهد. برای گروه‌بندی مناطق هم‌شکل در هر خوشه از روش خوشه‌یابی سلسله مراتبی استفاده می‌کنیم و از آنجایی که ما به دنبال یافتن گروه‌های هستیم که اعضای آن به اندازه کافی شباهت داشته باشند، برای تعیین فاصله بین دو خوشه از کوتاه‌ترین فاصله بین نمونه‌های آن دو استفاده می‌کنیم.



شکل ۱۲: مناطق شاخص زیر- کلمات عضو خوشه شکل ۱۱. این مناطق برای سه مقیاس ۱/۴، ۱/۲ و ۱ نمایش داده شده است.



شکل ۱۳: دندروگرام به دست آمده برای مناطق شاخص زیر- کلمات شکل ۱۲.

شکل ۱۳ دندروگرام شکل ۱۲ را نمایش می‌دهد که این نمودار برای مناطق محلی با مقیاس ۱/۴ رسم شده است. برای تشکیل گروه‌ها با استفاده از این نمودار درختی، آستانه‌ای برای برش انتخاب می‌شود و این آستانه برش بر اساس حداکثر فاصله قابل قبول بین نمونه‌های یک گروه تعیین شده و مقدار آن ۰/۵ در نظر گرفته شده است. گروه‌های مختلف در شکل ۱۳ مشخص شده و میانگین نمونه‌ها در هر گروه به عنوان نماینده آن گروه در نظر گرفته می‌شود. مجموعه این نماینده‌ها به عنوان مناطق شاخص خوشه در نظر گرفته می‌شود.

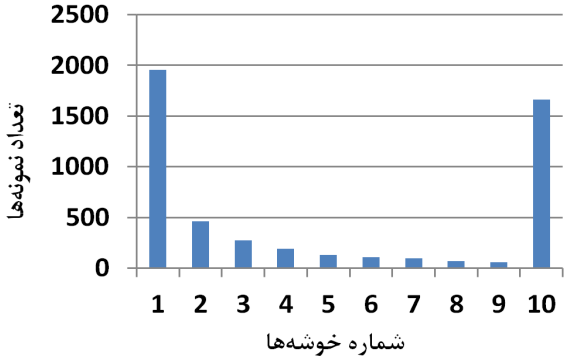
٦- انتخاب نهایی زیر- کلمات

بعد از طبقه‌بندی زیر- کلمه ورودی، فهرستی از خوشه‌های نزدیک‌تر به آن ایجاد می‌شود و اطمینان به نزدیک‌ترین خوشه در مرحله تأیید بررسی می‌شود.

در مرحله تأیید، بردارهای ویژگی محلی از شکل زیر- کلمه ورودی استخراج شده و با مناطق شاخص آن خوشه مقایسه می‌شوند که این مقایسه با محاسبه فاصله اقلیدسی بین هر بردار ویژگی و هر کدام از مناطق شاخص خوشه انجام می‌شود. اگر حداقل یکی از فاصله‌های محاسبه‌شده کمتر از آستانه از پیش تعیین شده باشد، این خوشه تأیید می‌شود و اعضای این خوشه به عنوان زیر- کلمات نهایی انتخاب می‌شوند و در غیر این صورت، الگوریتم برای خوشه بعد اجرا می‌شود. بررسی خوشه‌ها تا جایی ادامه می‌یابد که شرایط آستانه برآورده شود. در این صورت، این خوشه به همراه تمام خوشه‌های قبلی به عنوان پاسخ نهایی سامانه انتخاب می‌شوند. برای مثال اگر شرایط لازم برای چهارمین خوشه برآورده شود، خوشه‌های اول تا چهارم به عنوان خوشه‌های نهایی تعیین می‌شوند و دامنه جستجوی زیر- کلمه ورودی به جای ۱۰ خوشه اولیه به اعضای این ۴ خوشه محدود می‌شود و به این ترتیب، کاهش اندازه دیکشنری به اطمینان نتیجه طبقه‌بندی وابسته است. مقدار آستانه برای فاصله بردارهای ویژگی محلی و بخش‌های متمایزکننده هر خوشه، ۰/۵ تعیین شده است.

۷- آزمایش‌ها و تحلیل نتایج

مجموعه داده آموزشی که در این مقاله استفاده می‌شود، تصاویر بدنه ۶۸۹۵ زیر-کلمه است که با قلم لوتوس ۱۴ نگارش و چاپ شده و با درجه تفکیک ۴۰۰ نقطه در اینچ روبش شده‌اند (بخشی از مجموعه داده [۱۲]). برای انجام آزمایش‌ها و بررسی کارایی روش پیشنهادی، تعدادی نمونه آزمایشی نیز تولید شده که برای ایجاد نمونه‌های آزمایشی، ۱۰۰۰ زیر- کلمه از میان زیر-کلمات متداول فارسی [۱۲] به تصادف انتخاب شده است. این زیر- کلمات را با قلم لوتوس و در سه اندازه قلم ۱۰، ۱۲ و ۱۴ چاپ و با درجات تفکیک ۲۰۰، ۳۰۰ و ۴۰۰ نقطه بر اینچ روبش کرده‌ایم. با حذف نقاط و علایم از تصاویر زیر- کلمات، در مجموع ۵۰۰۰ تصویر به عنوان نمونه‌های آزمایشی به دست آمده است. برای بررسی روش پیشنهادی، آزمایش‌های مختلفی انجام شده و نتایج ارزیابی شده است.



شکل ۱۴: نمودار فراوانی تعداد خوشه‌های مناسب به دست آمده برای نمونه آزمایش.

شکل ۱۵: دقت طبقه‌بندی در بازه‌های مختلف شکل ۱۴.

جدول ۱: دقت طبقه‌بندی زیر– کلمات مجموعه آزمایش به ۳۰۰ خوشه بر اساس ویژگی‌های مکان مشخصه.

انتخاب دهم	انتخاب نهم	انتخاب هشتم	انتخاب هفتم	انتخاب ششم	انتخاب پنجم	انتخاب چهارم	انتخاب سوم	انتخاب دوم	انتخاب اول
۹۹/۳۴	۹۹/۲۴	۹۸/۵۶	۹۷/۹۶	۹۷/۲۲	۹۶/۶	۹۵/۴۲	۹۳/۸۳	۹۰/۴۲	۷۷/۷۲

جدول ۲: دقت طبقه‌بندی زیر– کلمات بر اساس مناطق متمایزکننده هر خوشه در سه مقیاس ۱/۴، ۱/۲ و ۱.

	مقیاس ۱/۴	مقیاس ۱/۲	مقیاس ۱
دقت طبقه‌بندی (٪)	۹۷/۳۴	۹۹/۱۷	۹۴/۶۸
میانگین تعداد خوشه‌های نهایی	۶/۴۳	۴/۸	۴/۶۲

۷–۱ تأثیر روش پیشنهادی در کاهش اندازه فضای جستجو

با استفاده از روش پیشنهادی این مقاله، برای کاهش اندازه دیکشنری، تعداد خوشه‌هایی که به عنوان خوشه‌های نهایی انتخاب می‌شوند برای زیر– کلمات مختلف متفاوت است و این تعداد می‌تواند از یک تا ۱۰ خوشه متغیر باشد. انتخاب تنها یک خوشه به عنوان خوشه نهایی، مربوط به حالتی است که اولین خوشه تأییدشده باشد و در این حالت فضای جستجو به بیشترین اندازه ممکن کاهش داده شده است. انتخاب هر ۱۰ خوشه هم مربوط به زمانی است که بعد از بررسی هر ده خوشه، هیچ کدام تأیید نشده باشند که در این حالت، اعضای هر ۱۰ خوشه به عنوان خوشه‌های نهایی در نظر گرفته می‌شود و استفاده از روش پیشنهادی تأثیری در کاهش بیشتر فضای جستجو ندارد. شکل ۱۴ نمودار فراوانی تعداد خوشه‌های نهایی را برای مجموعه ۵۰۰۰ نمونه آزمایشی نشان می‌دهد.

با توجه به شکل برای حدود ۴۰٪ از نمونه‌های آزمایشی، اولین خوشه به عنوان خوشه نهایی تشخیص داده شده و فراوانی زیر– کلمات با افزایش شمار خوشه‌های نهایی به تدریج کاهش می‌یابد. این کاهش بیانگر آن است که هر چقدر از نتایج اول طبقه‌بندی فاصله بگیریم، احتمال تأییدشدن خوشه‌ها نیز کاهش می‌یابد. مقدار نمودار در خوشه دهم، تعداد زیر– کلماتی را نشان می‌دهد که در بررسی ۱۰ خوشه انتخابی، هیچ کدام تأیید نشده‌اند و کل ۱۰ خوشه به عنوان خوشه نهایی در نظر گرفته شده است.

میانگین تعداد خوشه‌هایی که به عنوان خوشه‌های نهایی برای این ۵۰۰۰ نمونه انتخاب شده است، ۴/۸ خوشه است. این عدد را می‌توان به این صورت تعبیر کرد که با استفاده از روش پیشنهادی، تعداد خوشه‌های نهایی از ۱۰ خوشه به حدود ۵ خوشه کاهش خواهد یافت. با توجه به آنچه در بخش ۴–۲ بیان شد، کمینه و بیشینه تعداد اعضای هر خوشه ۱ و ۵۴ نمونه است و به این ترتیب فضای جستجو برای مراحل بعدی به شکل قابل ملاحظه‌ای کاهش می‌یابد.

البته این نتیجه‌گیری تنها در شرایطی صحیح است که دقت نیز حفظ شده باشد. شکل ۱۵ دقت سامانه را در هر کدام از ستون‌های هیستوگرام شکل ۱۴ نشان می‌دهد. به عنوان مثال، دقت طبقه‌بندی در نمونه‌هایی که اولین خوشه به عنوان خوشه نهایی آنها انتخاب شده، نزدیک به ۱۰۰٪ است و بنابراین در تمام مواردی که اولین خوشه به عنوان خوشه نهایی تأیید شده، این تصمیم به درستی گرفته شده و زیر– کلمه ورودی عضو آن خوشه است. همان طور که پیش از این اشاره شد، این نمونه‌ها حدود ۴۰٪ کل نمونه‌های آموزشی را تشکیل می‌دهند و دقت طبقه‌بندی برای نمونه‌هایی که خوشه دوم در آنها تأیید شده، مقدار کمتری است. کاهش میزان دقت برای تعداد بالاتر خوشه‌ها هم ادامه پیدا می‌کند و این کاهش بدین معنی است که با افزایش فاصله یک خوشه، احتمال وقوع خطا در تأیید آن نیز بالا می‌رود. افزایش دقت برای ۸ و ۹ خوشه به دلیل تعداد کم نمونه‌هایی است که در این بازه قرار دارند که این نمونه‌ها، نمونه‌هایی هستند که خوشه‌های هشتم یا نهم آنها تأیید شده است. احتمال تأیید خوشه‌های دورتر به دلیل فاصله زیاد از نمونه آزمایشی، بسیار پایین است.

در شکل ۱۵ دقت سامانه به تفکیک تعداد خوشه‌های نهایی نمایش داده شده است. برای مقایسه جامع‌تر، دقت روی کل نمونه‌های آزمایشی نیز محاسبه شده که این دقت ۹۹/۱۷٪ به دست آمده است. بدون استفاده از روش تأیید پیشنهادی، برای رسیدن به این دقت نیاز است که حداقل ۹ خوشه نزدیک‌تر به زیر– کلمه ورودی انتخاب شود (مقادیر دقت طبقه‌بندی به ۳۰۰ خوشه، بر اساس ویژگی‌های مکان مشخصه در جدول ۱ آمده است). این در حالی است که با استفاده از روش تأیید به طور میانگین با انتخاب ۴/۸ خوشه برای هر نمونه ورودی می‌توان به همین دقت رسید. امکان افزایش مقدار این دقت با تغییر مقادیر پارامترها وجود دارد. پر واضح است که در این شرایط، میانگین تعداد خوشه‌های نهایی نیز افزایش خواهد یافت.

۷–۲ بررسی تأثیر اندازه مناطق محلی

برای استخراج مناطق محلی از شکل زیر– کلمات، قاب‌هایی در سه مقیاس ۱/۴، ۱/۲ و ۱ برابر اندازه مستطیل محیطی‌شکل زیر– کلمه استفاده شده و نتایج حاصل از طبقه‌بندی زیر– کلمات به ازای هر سه حالت در جدول ۲ آمده و با توجه به این جدول، با استفاده از مقیاس ۱/۲ دقت نسبت به دو حالت دیگر بیشینه است. میانگین تعداد خوشه‌های

شکل ۱۶: تصاویر نمونه‌هایی از خطای طبقه‌بندی در مقیاس ۱/۲، طبقه‌بندی این زیر- کلمات دست کم در یکی از مقیاس‌های ۱/۴ یا ۱ به درستی انجام شده است.

انتخابی نیز در این مقیاس نسبت به مقیاس ۱/۴ کمتر است. با وجود آن که میانگین تعداد خوشه‌ها در مقیاس ۱ کمی کمتر از مقیاس ۱/۲ گزارش شده است اما دقت طبقه‌بندی در این مقیاس مناسب نیست.

برای بررسی بیشتر، تصویر تعدادی از زیر- کلمات که با مقیاس ۱/۲ به خطا طبقه‌بندی شده‌اند در شکل ۱۶ آمده است که تمام این زیر- کلمات دست کم در یکی از مقیاس‌های ۱/۴ یا ۱ به درستی طبقه‌بندی شده‌اند.

با توجه به این شکل در بین نمونه‌های خطا، زیر- کلمات با شکلی مشابه وجود دارند. به عنوان مثال، ۵ زیر- کلمه‌ای که در اولین ردیف نمایش داده شده‌اند، همگی عضو یک خوشه هستند. وقوع خطا به ازای نمونه‌های متعدد از یک خوشه ممکن است از آنجا ناشی شود که منطقه‌بندی با مقیاس ۱/۲ برای همه خوشه‌ها مناسب نیست و از این رو می‌توان برای تعیین مناطق شاخص منتسب به هر خوشه، از مقیاس مناسب آن خوشه استفاده کرد. به این ترتیب هر نمونه ورودی با مناطق شاخص هر خوشه در مقیاس مناسب آن خوشه مقایسه می‌شود. بررسی جزئی‌تر تأثیر اندازه مقیاس از حوصله این مقاله خارج است و نیاز به تحقیق جامع‌تری دارد.

۷-۳ بررسی نمونه‌های خطا

در شکل ۱۷ تصاویر تعدادی از زیر- کلمات نشان داده شده که با استفاده از روش پیشنهادی به خطا طبقه‌بندی شده‌اند و زیر- کلمات بر اساس تعداد خوشه‌های نهایی متناظر مرتب شده‌اند. به این ترتیب برای زیر- کلمات ردیف rام، r خوشه اول به عنوان خوشه نهایی انتخاب شده است اما این نمونه‌ها عضو این r خوشه نهایی نبوده‌اند. خانه‌های خالی مربوط به مواقعی است که تعداد خطا کمتر از ۵ بوده است. با بررسی این نمونه‌ها، عوامل مؤثر ایجاد خطا را می‌توان به موارد زیر تقسیم کرد:

۱) بسیاری از زیر- کلمات فاقد بالارونده یا پایین‌رونده هستند. با توجه به این که روش مکان مشخصه در تشخیص بالا و پایین‌رونده‌ها توانمندتر است، عدم وجود این اجزا دقت مرحله طبقه‌بندی بر اساس ویژگی‌های سراسری را کاهش می‌دهد. این امر، احتمال وقوع خطا در کل سامانه را بالاتر می‌برد.

۲) چسبیدگی سرکش حرف "گ" به بدنه در تعدادی از نمونه‌ها دیده می‌شود و با کوچک‌شدن اندازه قلم و کاهش درجه تفکیک تصویربرداری، احتمال چسبیدگی نقاط یا علایم به بدنه کلمات افزایش می‌یابد.

۳) همان طور که در بخش ۷-۲ نیز اشاره شد، وجود زیر- کلمات با شکل مشابه در این مجموعه بیانگر این است که استخراج مناطق محلی با اندازه ثابت قاب برای تمام نمونه‌ها مناسب نیست. انتخاب مناسب ابعاد مناطق محلی با توجه به طول متغیر کلمات می‌تواند احتمال خطا را کاهش دهد.

تعداد خوشه	انتخاب اول	انتخاب دوم	انتخاب سوم	انتخاب چهارم	انتخاب پنجم
۱	حفد	بعد	بعف	صف	بص
۲	بک	حب	لعبی	مصمص	حسمس
۳	سبی	مبی	بعو	بک	بمحا
٤	مصمص	فعه	بسه	بمحا	بمحا
٥	بمسه	مبی	بمحا	بک	ـــــ
٦	بعب	بک	بگبد	گبا	ـــــ
۷	مصد	بگ	بگ	گلص	گلص
۸	گلص	ـــــ	ـــــ	ـــــ	ـــــ
۹	ـــــ	ـــــ	ـــــ	ـــــ	ـــــ
۱۰	فب	بمب	بعب	معبه	بها

شکل ۱۷: نمونه‌هایی از خطای روش پیشنهادی در طبقه‌بندی. نمونه‌ها بر حسب تعداد خوشه‌هایی مرتب شده‌اند که در مرحله تصمیم‌گیری نهایی به دست آمده‌اند.

۸- نتیجه‌گیری

با توجه به پیوسته‌نویسی در خط فارسی، کارایی روش‌های مبتنی بر شکل کلی در توصیف زیر- کلمات انکارناپذیر است. در بازشناسی کلمات بر اساس شکل کلی، تعداد کلاس‌ها با تعداد کلمات مورد بررسی برابر است، از این رو نیازمند به کارگیری روش‌هایی برای کاهش دامنه جستجو هستیم. با خوشه‌بندی مجموعه کلمات، فضای جستجو به بخش‌هایی افراز می‌شود که هر کدام تعدادی از زیر- کلمات را در خود جای می‌دهد و با انتخاب خوشه متناظر با زیر- کلمه ورودی، فضای جستجو به اعضای آن خوشه محدود می‌شود. انتخاب دقیق خوشه متناظر با ورودی، علاوه بر این که دقت سامانه نهایی را افزایش می‌دهد، می‌تواند فضای جستجو را نیز محدودتر کند. در این مقاله روشی را برای کاهش اندازه دیکشنری ارائه کردیم که با دقت بالایی به تعیین خوشه‌های مناسب می‌پردازد. در روش ارائه‌شده، برای هر خوشه مناطقی تعیین می‌شود که اعضای آن خوشه را از سایر خوشه‌ها متمایز می‌کند و این مناطق- که مناطق شاخص نامیده می‌شوند- برای تصمیم‌گیری در مورد تعلق نمونه ورودی به هر خوشه به کار گرفته می‌شود.

مجموعه کل زیر- کلمات، ابتدا بر اساس ویژگی‌های سراسری شکل به ۳۰۰ خوشه تقسیم می‌شود و هر زیر- کلمه ورودی به این خوشه‌ها طبقه‌بندی شده و ۱۰ خوشه نزدیک‌تر انتخاب می‌شود. سپس این خوشه‌ها در مرحله تأیید بررسی شده و خوشه‌های نهایی از بین آنها انتخاب می‌شوند. تأیید هر خوشه بر میزان تطابق زیر- کلمه ورودی با خصوصیات متمایزکننده هر خوشه استوار است. مناطقی از شکل کلمات که بین نمونه‌های یک خوشه بیشترین شباهت را دارند و در عین حال بیشترین تفاوت را با نمونه‌های سایر خوشه‌ها ایجاد می‌کند، به عنوان مناطق شاخص آن خوشه انتخاب می‌شوند. در مرحله تأیید، مناطق محلی از شکل زیر- کلمه ورودی استخراج شده و با مناطق شاخص خوشه مقایسه می‌شود. نتایج مرحله تأیید برای تصمیم‌گیری و انتخاب خوشه‌های نهایی به کار گرفته می‌شود. روش ارائه‌شده بر مجموعه تصاویر ۶۸۹۵ زیر- کلمه فارسی اجرا شده و از نتایج طبقه‌بندی ۵۰۰۰ نمونه آزمایش برای ارزیابی روش پیشنهادی استفاده شد.

آزمایش‌ها مؤید کارایی قابل توجه روش پیشنهادی در کاهش فضای جستجوی زیر- کلمات است. علاوه بر آن چه در بخش ۷ در مورد نتایج آزمایش‌های صورت‌گرفته بر روی نمونه‌هایی از یک قلم ارائه شد، آزمایش‌های دیگری را نیز برای تحقیق عملکرد این روش در کاربردهایی با تعداد بیش از یک قلم انجام دادیم. تنوع شکل زیر- کلمات در قلم‌های مختلف باعث می‌شود تعداد مناطق شاخصی که برای گروه‌های مختلف کلمات به دست خواهد آمد، کمتر شود. با اضافه‌شدن قلم‌هایی مثل قلم "هما" که از نظر شکل اجزا تفاوت بیشتری با قلم‌های معمول دارند، این قضیه بیشتر نمایان می‌شود. کاهش تعداد مناطق شاخص، امکان تأیید خوشه‌های انتخابی اولیه را کاهش می‌دهد و به این ترتیب فضای جستجو به میزان کمتری کاهش خواهد یافت.

علاوه بر افزایش تعداد قلم‌ها، سایر عواملی که موجب تنوع شکل یک زیر- کلمه می‌شوند نیز کارایی روش ما را کاهش می‌دهد که وجود نویز در تصاویر و تفاوت‌های نگارشی در متون دست‌نویس از این جمله هستند. برای ارتقای کارایی روش پیشنهادی در چینی مواردی، استفاده از مدل‌های مبتنی بر احتمال می‌تواند در تحقیقات به کار گرفته شود. رویکرد دیگری که می‌تواند در تحقیقات آتی جهت بهبود عملکرد سیستم بررسی شود، بهبود روش امتیازدهی به مناطق مختلف شکل است. در روش پیشنهادی این مقاله، امتیاز هر منطقه تنها بر اساس ویژگی‌های شکل آن تعیین می‌شود حال آن که می‌توان از اطلاعات مکانی هر منطقه نیز استفاده کرد. به این ترتیب علاوه بر ویژگی‌های شکلی، موقعیت قرارگیری هر منطقه در کل شکل زیر- کلمه نیز در نظر گرفته می‌شود.

مراجع

[1] T. Adamek, N. E. Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents," *Int. J. of Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 153-165, 2007.

[2] J. R. Pinales, R. J. Rivas, and M. J. C. Bleda, "Holistic cursive word recognition based on perceptual features," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1600-1609, 1 Oct. 2007.

[3] A. Amin, "Recognition of printed arabic text based on global features and decision tree learning techniques," *Pattern Recognition*, vol. 33, no. 8, pp. 1309-1323, 2000.

[4] A. Ebrahimi and E. Kabir, "A pictorial dictionary for printed farsi sub-words," *Pattern Recognition Letters*, vol. 29, no. 5, pp. 656-663, 2008.

[5] K. Zagoris, K. Ergina, and N. Papamarkos, "A document image retrieval system," *Engineering Application of Artificial Intelligence*, vol. 23, no. 6, pp. 872-879, 2010.

[6] S. Bai, L. Li, and C. L. Tan, "Keyword spotting in document images through word shape coding," in *Proc. 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09*, pp. 331-335, 26-29 Jul. 2009.

[7] L. Li, S. Lu, and C. L. Tan, "A fast keyword-spotting technique," in *Proc. 9th Int. Conf. on Document Analysis and Recognition*, *ICDAR'07*, pp.68-72, 23-26 Sep. 2007.

[8] S. Lu and C. L. Tan, "Document image retrieval through word shape coding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1913-1918, Nov. 2008.

[9] J. A. Rodriguez-Serrano and F. Perronnin, "Handwritten word - spotting using hidden markov models and vocabularies," *Pattern Recognition*, vol. 42, no. 9, pp. 2106-2116, Sep. 2009.

[10] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *Int. J. on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 139-152, Apr. 2007.

[11] Y. Lu and C. L. Tan, "Information retrieval in document image databases," *IEEE Trans. on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1398-1410, Nov. 2004.

[۱۲] ا. ابراهیمی، *استفاده از شکل کلی زیر-کلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی*، رساله دکتری مهندسی برق-الکترونیک، دانشگاه تربیت مدرس، تهران، ۱۳۸۴.

[۱۳] ح. خسروی و ا. کبیر، "ارزیابی روش‌های بازشناسی متون فارسی بر مبنای شکل کلی زیر- کلمات،" *نشریه مهندسی برق و کامپیوتر ایران*، جلد ۷، شماره ۴، صص. ۲۸۰-۲۶۷، زمستان ۱۳۸۸.

[14] S. Madhvanath, G. Kim, and V. Govindaraju, "Chain code contour processing for handwritten word recognition," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 21, no. 9, pp. 928-932, Sep. 1999.

[۱۵] ا. ابراهیمی و ا. کبیر، "یک روش دومرحله‌ای برای بازشناسی زیر- کلمات چاپی،" *نشریه مهندسی برق و کامپیوتر ایران*، جلد ۲، شماره ۲، صص. ۶۲-۵۷، پاییز و زمستان ۱۳۸۳.

[16] S. G. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 149-164, Feb. 2001.

[17] A. Rehman and T. Saba, "Off-line cursive script recognition: current advances, comparisons and remaining problems," *Artificial Intelligence Review*, vol. 37, no. 4, pp. 261-288, 2012.

[18] L. M. Lorigo and V. Govindaraju, "Off - line arabic handwriting recognition: a survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 712-724, May 2008.

[۱۹] م. ش. شهرضا، *تشخیص کلمات و ارقام دست‌نویس فارسی به وسیله شبکه‌های عصبی (خط نسخ)*، رساله دکتری مهندسی برق و کامپیوتر، دانشگاه امیرکبیر، تهران، ۱۳۷۴.

[۲۰] ر. عزمی، *بازشناسی متون چاپی فارسی*، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، تهران، ۱۳۷۸.

[21] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Handwritten farsi (arabic) word recognition: a holistic approach using discrete HMM," *Pattern Recognition*, vol. 34, no. 5, pp. 1057-1065, 2001.

[22] M. H. Shirali-Shahreza, K. Faez, and A. Khotanzad, "Recognition of handwritten farsi numerals by zernike moments features and a set of class-specific neural network classifiers," in *Proc. on Int. Conf. of Signal Processing Applications and Technology*, pp. 998-1003, 18-20 Oct. 1994.

[23] P. Shilane and T. Funkhouser, "Distinctive regions of 3D surfaces," *ACM Trans. on Graphics*, vol. 26, no. 2, Article 7, Jun. 2007.

[24] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of 4th Alvey Vision Conf.*, pp. 147-151, 1988.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 886-893, 2005.

[۲۶] ه. داودی و ا. کبیر، "تعیین بخش‌های مهم در شکل زیر-کلمات چاپی،" *بیستمین کنفرانس مهندسی برق ایران*، ۲۴۴۷-۲۴۴۲ صص.، تهران، ۲۸-۲۶ اردیبهشت ۱۳۹۱.

هما داودی کارشناسی و کارشناسی ارشد مهندسی برق را به ترتیب در سال ۸۶ از دانشگاه تبریز و سال ۸۸ از دانشگاه تربیت مدرس دریافت کرد. او هم اکنون مشغول تحصیل در مقطع دکترا در دانشگاه تربیت مدرس است. زمینه‌های تحقیقاتی مورد علاقه او، بازشناسی الگو، بینایی ماشین و پردازش تصویر است.

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.