نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

استخراج بهترین ویژگی از متون فارسی با استفاده از تجزیه و تحلیل مؤلفه‌های اصلی با کمک میانگین یادآوری و الگوریتم ژنتیک

حمید حسن‌پور[۱]، علی قنبری سرخی[۲]، اشکان پارسی[۳]
[۱] دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات، شاهرود h_hassanpour@yahoo.com
[۲] دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات، شاهرود ali.ghanbari289@gmail.com
[۳] دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات، شاهرود a.parsi@shahroodut.ac.ir

چکیده

طبقه بندی و استخراج ویژگی متون فارسی به دلیل وجود ویژگی‌های بسیار، تکراری و بی اهمیت، فرایندی بسیار سخت و پیچیده خواهد بود. از آنجا که این موضوع به صورت محدود مورد مطالعات قرار گرفته است، هدف از مقاله حاضر، استخراج بهترین ویژگی‌های متن فارسی با استفاده از تجزیه و تحلیل مؤلفه‌های اصلی[۱] (PCA) با کمک معیار میانگین یادآوری و الگوریتم ژنتیک خواهد بود. این مطالعه با در اختیار داشتن مجموعه داده‌های استاندارد روزنامه همشهری که در پنج طبقه تقسیم شده بودند، انجام شد. با استفاده از روش وزن دهی ویژگی مبتنی بر اطلاعات کلاس در حوزه طبقه‌بندی مستندات[۲] (TFCRF) و روش‌های طبقه‌بندی نزدیکترین همسایه[۳] (KNN) و بیزین[۴] در روش پیشنهادی، نتایج به دست آمده نشان داد که دقت طبقه بندی متون فارسی به صورت قابل توجهی افزایش و مدت زمان تست با ویژگی‌های استخراج شده با روش پیشنهادی کاهش خواهد یافت.

کلمات کلیدی

طبقه‌بندی متون، وزن‌دهی ویژگی، تجزیه وتحلیل مؤلفه‌های اصلی، میانگین یادآوری، الگوریتم ژنتیک.

۱- مقدمه

افزایش روزافزون دامنه اطلاعات به قدری است که عمل دسته-بندی آنها را امری ضروری نموده است؛ از تارنماهای بزرگ اطلاع‌رسانی تا مراجع عظیم علمی با سرعتی باور نکردنی در حال رشد بوده و نیاز به دسته‌بندی اطلاعات و قرارداد آنها در زیر مجموعه‌های گوناگون به شدت احساس می‌شود. با توجه به این توضیح، دسته‌بندی متون و انتساب اسناد به دسته‌هایی مشخص و از پیش تعیین شده، در دهه اخیر توجه بسیاری را به خود جلب نموده است. طبقه‌بندی اسناد، در دنیایی پیشرفته در کنار ماشین‌هایی با قدرت پردازش بالا، از طریق استفاده از عنصر یادگیری در این ماشین‌ها، بوسیله نمونه‌های داده‌ای موسوم به نمونه‌های آموزش و آزمون انجام می‌گیرد، تا فرآیند طبقه‌بندی را به خوبی انسان انجام دهد. از میان روش‌هایی که جهت دسته‌بندی متون استفاده می‌شوند می‌توان به روش‌های نزدیک‌ترین همسایگی [۲-۱]، درخت تصمیم‌گیری [۳]، یادگیری آماری (نظیر مدل‌های رگرسیون [۴] و مدل بیزین [۵])، شبکه های عصبی [۶]، ماشین های برداری پشتیبان [۷] و غیره اشاره کرد.

اکثر سیستم‌هاي طبقه‌بندي متون، جهت طبقه‌بندی متون انگليسي طراحي شده‌اند و تحقیقات کمی بر روی دسته‌بندی متون فارسي انجام گرفته است. مقاله [۸]، شش روش بازیابی اطلاعات را با استفاده از مجموعه متنی همشهری ارزیابی می‌کند. همچنین راهکاری برای ترکیب این روش‌ها جهت افزایش کیفیت جواب‌های سامانه مورد تحقیق، ذکر شده است. در این مقاله از دو روش فضای بردار و چهار روش مدل کردن زبان بهره گرفته شده است.

برای دسته‌بندی متون فارسی در مرجع [۹] با استفاده از روش KNN و در نظر گرفتن دو حالت با حذف کلمات زائد و بدون حذف آن‌ها به دسته‌بندی متون فارسی پرداخته که حذف کلمات زائد نشان دهنده اندکی بهبود در نتایج بدست آمده، بوده است. در [۱۰] به بررسی دسته‌بندی متن فارسی با استفاده از الگوریتم KNN و FKNN (fuzzy KNN) پرداخته شده و هدف از مقاله بررسی و مقایسه دو الگوریتم مذکور برای دسته‌بندی متن فارسی و ترکیب آنها با روش‌های انتخاب ویژگی بهره اطلاعات[۵] (IG) و فرکانس سند[۶] (DF) است.

یکی از مشکلات اساسی در دسته‌بندی متون حجم بالای ویژگی‌های استخراج شده می‌باشد، که در بسیاری از الگوریتم-های موجود، این حجم بالا منجر به کندی طبقه‌بندی کننده و