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

بهبود دسته بندي متون فارسي در روش همسايگی وزن دار

فرزین یغمایی[۱]، سعید تعبدی[۲]
[۱]f_yaghmaee@semnan.ac.ir
[۲]s.taabbody@gmail.com
[۱]دانشگاه سمنان، دانشکده مهندسی برق و کامپیوتر

چکیده:

با رشد روز افزون منابع اطلاعاتي و حجم مقالات و مطالب توليد شده در زمينه‌هاي مختلف و به شكل‌هاي متنوع اعم ازرسانه های مختلف ديجيتال نياز به دسترسي آسان اطلاعات نيز افزايش مي‌يابد. يكي از نياز هاي اوليه در بالا بردن سرعت دسترسي به اطلاعات و پردازش اين مطالب كه غالباً داراي حجم بالايي نيز مي‌باشند، دسته بندي[۱] اين اطلاعات در طبقات مختلف مي‌باشد. دسته بندي متون به عمل برچسب زدن[۲] يا تفكيک يک متن در قالب يكي از دسته هاي از پيش تعيين شده گفته مي‌شود. در اين مقاله به بررسي عملكرد الگوريتم WKNN[۳] با استفاده از معيار وزن‌دهي tf-idf مي‌پردازيم. همچنين براي بالابردن دقت در انتخاب طبقه صحيح و به منظور افزايش كارايي الگوريتم از روش ميانگين‌گيري از داده‌ها به عنوان معيار ارزيابي استفاده ميكنيم. نتايج به دست آمده از تفكيک متون فارسي با استفاده از روش‌هاي فوق نشان دهنده دقت ۸۹ درصد مي‌باشد.

كلمات كليدي:
دسته بندي متن، تفكيک كننده WKNN، نزديكترين همسايگي، وزن‌دهي

۱- مقدمه

با پيشرفت در علوم كامپيوتري و گسترش كاربرد آن در تمام زمينه‌ها، حجم ذخيره و پردازش اطلاعات افزايش يافته و امكان دسترسي به اين اطلاعات از محل‌هاي ديگر نيز با استفاده از شبكه هاي موجود بين كامپيوتر ها امكان‌پذير شده است. به منظور بهبود نحوه دسترسي به اين اطلاعات، وجود سيستم‌هايي براي انجام اعمالي از قبيل فيلترينگ، طبقه بندي، جستجو و ... ضروري مي‌باشد.

دسته بندي متن فرايند تصميم گيري براي نسبت دادن يک متن به يک گروه خاص از متون مي باشد.از اين سيستم براي دسته بندي متن ها يا سايت ها براي دسترسي آسان به اطلاعات آنها استفاده مي شود. به منظور دسته بندي متون مي توان از الگوريتم هاي متنوعي استفاده نمود كه از آن جمله مي توان به شبكه هاي عصبي، نزديكترين همسايگي[۴]، SVM[۵] ، Naive Bayes و غيره اشاره نمود. موارد ذكر شده جزء الگوريتم هاي پركاربرد در حوزه IR[۶] مي‌باشند[۱][۲][3]

در زمينه دسته بندي متون انگليسي روشها و مقالات متنوعي وجود دارد كه هر كدام ويژگي ها خاص خود را دارا مي-باشند اما در زمينه دسته بندي متون فارسي كار چنداني صورت نگرفته است.

در اين مقاله، به منظور دسته‌بندي خودكار متون فارسي از روش شاخص گذاري متون، تعيين ميانگين وزنها و همچنين از تفكيک كننده موسوم به K مين همسايه نزديک استفاده شده است.

يكي از راههاي موثر در دسته‌بندي متون امكان ايجاد رابطه-اي بين كلمات موجود در متن و گروه مرتبط با آن متن است كه در صورت يافتن چنين رابطه‌اي پيش بيني گروه مورد نظر امكان پذير خواهد بود. برخي از تحقيقات انجام شده بر روي دسته بندي متون فارسي عبارتند از: منبع [2] كه در آن از ۳ روش براي طبقه بندي متون فارسي استفاده شده است و همچنين منبع [1]