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

خوشه بندی متون مبتنی بر مرکز دسته با استفاده از روش SVD و بهره گیری از نقاط همسایگی

محمدرضا علاقه بند[۱] ؛ محمد رضا سعیدی محمدی[۲] ؛ میر حسین دزفولیان[۳]
mr.alagheband@basu.ac.ir[۱]
m.saeedi@basu.ac.ir[۲]
dezfoulian@basu.ac.ir[۳]

چکیده

دسته بندی متن، یکی از شاخه هایی است که امروزه توجه گسترده ای را به خود معطوف داشته است که در آن متن هاي با محتويات مشابه به يك دسته اختصاص داده مي‌شوند . در این مقاله روشی ترکیبی از خوشه بندی مبتنی بر مرکز دسته ها، K-means ، به همراه روش SVD مورد مطالعه قرار گرفته که عمل خوشه بندی در آن، با معرفی دو مفهوم جدید نقاط همسایگی و ارتباط ، انجام پذیرفته است . استفاده از SVD مي تواند با حفظ ساختار زیربنایی در فضای ویژگی، اثر نویز را کاهش دهد. با استفاده از مفاهیم ارائه شده ، روشی جهت انتخاب مراکز اولیه دسته ها و رابطه ای جهت محاسبه مشابهت بردار متن ها با این مفاهیم معرفی می‌گردد. در انتها سه آزمایش مختلف بر روی یک مجموعه از پنجاه و شش متن انجام گرفته‌است. در آزمایش ابتدایي با استفاده از الگوریتم مقدماتي K-means به نتیجه ۵۷,۱۴ درصد سپس در ازمایش دوم با اعمال مفاهیم همسایگي و پیوند به نتیجه ۸۷,۵ درصد و آزمایش سوم با استفاده ترکیبي از مفاهیم ارائه شده در آزمایش دوم و الگوریتم SVD ، دسته بندي متون را به ۹۲,۸۶ درصد بهبود می‌بخشد.

کلمات کلیدی

خوشه بندی، K-means ، SVD ، نقاط همسایگی، مفهوم ارتباط

۱- مقدمه

رشد حجم مستندات دیجیتالي در سال هاي اخير ، مديريت اطلاعات را با مشكلات عديده اي روبرو ساخته است كه لازم است روش هاي دقيقي جهت گروه بندی این حجم از اطلاعات ارائه گردد. به همین علت تا به حال تحقیقات بسیار زیادی در زمینه خوشه بندی در دنیای بازیابی اطلاعات انجام شده است [۱،۲] .در مقاله حاضر، روش k-means [3-6] به کار گرفته شده است، در ابتدا تعداد k نقطه را به عنوان نماینده هر خوشه در نظر گرفته و سپس هر متن بسته به میزان شباهت، به یکی از این خوشه ها منتصب می‌شود. سپس مجددا مرکز هر خوشه محاسبه می‌گردد. این روند تا جایی که مراکز دسته ها تغییر نامحسوس داشته و یا اصلا تغییر نکنند تکرار میگردد.

از زمان ارائه الگوریتم K-mean تا کنون نسخه های متفاوت ان در بسیاری از سیستم ها در کاربرد های مختلف استفاده شده است. از کاربردهای ان می‌توان به کاربرد های تجاری چون دسته بندی هوشمند سهام در بازار بورس[7] ، سیستم های پیشنهاد دهنده خرید کالا انلاین [8] ،کاربر د های صنعتی در دسته بندی انواع زغال سنگ [9] ، تشخیص وضعیت ترافیک در شهر [10] ، و یا کاربرد های پزشکی در دسته بندی بافت های مغز[11] اشاره کرد.در ادامه یک روش جهت انتخاب نقاط اولیه مراکز ارائه شده تا نماینده هایی مناسب به عنوان مرکز هر خوشه انتخاب شوند. از این جهت، مفهومی به نام امتیاز نقاط کاندید را معرفی کرده ایم.برای یافتن نقاط کاندید لازم است از سه معیار استفاده شود : میزان شباهت بین دو متن ، مقدار مفهوم ارتباط آنها و تعداد نقاط همسایگی آنها . در نهایت از مفهوم تجزیه مقادیر تکین (SVD)[۱] استفاده شده تا با بهره‌گیری از این مفاهیم خوشه بندی اسناد را بهبود بخشیم.

۲- خوشه بندی متون

در این بخش روند کار الگوریتم ارائه شده در این مقاله را معرفی می کنیم .

۲-۱- پیش پردازش

معمولا متن ها حاوی کلمه هایی هستند که معنی خاصی ندارند. از این دست لغات می توان به حروف اضافه اشاره نمود که بهتر است از متن ها حذف گردند. بعلاوه به جای استفاده از لغاتی که در متن ها وجود دارند بایستی هنگام بررسی شباهت متن ها