علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۶، شماره ۲، ۱۳۹۷
صفحات ۳۹- ۴۸
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

انتخاب هوشمندانه مراکز اولیه در الگوریتم خوشه‌بندی K-means به‌منظور بهبود تشخیص موضوع

سپهر آروین علی ورداسبی هشام فیلی آزاده شاکری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده

تشخیص موضوع یکی از مسائل حوزه‌ی پردازش زبان طبیعی است که در سال‌های اخیر همواره مورد توجه بوده و از زوایای متفاوتی مورد پژوهش قرارگرفته است. هدف کلی در این مسئله خوشه‌بندی اسناد متنی در دسته‌های مختلف است به‌گونه‌ای که اسناد موجود در هر خوشه موضوع یکسانی داشته باشد. بخش قابل‌توجهی از راه‌حل‌های ارائه شده برای این مسئله از الگوریتم‌های خوشه‌بندی مانند K-means استفاده می‌کنند. علاوه بر روش‌های مبتنی بر خوشه‌بندی اسناد، در دسته‌ای از پژوهش‌ها برای حل مسئله تشخیص موضوع از روش‌های مدل‌سازی موضوعی استفاده شده است.

در این پژوهش ابتدا حساسیت قابل توجه الگوریتم K-means به انتخاب مراکز اولیه به‌صورت عملی نشان داده می‌شود و سپس روشی برای انتخاب هوشمندانه مراکز اولیه ارائه می‌شود که استفاده از آن کیفیت الگوریتم K-means را در مسئله‌ی تشخیص موضوع ارتقاء می‌دهد. روش پیشنهاد شده برای تشخیص موضوع در این مقاله با بهره‌گیری از مدل‌سازی موضوعی LDA (Dirichlet Allocation Latent)، پس از انتخاب هوشمندانه مراکز اولیه، اقدام به خوشه‌بندی اسناد براساس موضوع آن‌ها می‌کند. در روش ارائه شده فاصله اسناد براساس توزیع موضوع حاصل از LDA آن‌ها محاسبه شده است. آزمایش‌ها نشان می‌دهند که استفاده از روش ارائه شده باعث بهبود چشم‌گیر کیفیت تشخیص موضوع نسبت به روش LDA در دو مجموعه از سه مجموعه دادگان مورد آزمایش می‌شود. همچنین در مقایسه با روش K-means++ برای انتخاب مراکز اولیه، در روش ارائه شده‌ی ما انتخاب مراکز اولیه در دو مجموعه دادگان همیشه مناسب‌تر بوده و احتمال بهتر بودن مراکز انتخابی در مجموعه دادگان دیگر مورد آزمایش برابر با ۷۰ درصد است.

کلمات کلیدی: LDA (Latent Dirichlet Allocation)، خوشه‌بندی، تعیین مراکز اولیه، K-Means، معیار فاصله، Silhouette.

۱- مقدمه

گسترش سریع وب و شبکه‌های اجتماعی موجب شده است که روزانه حجم زیادی از اطلاعات توسط خبرگزاری‌ها و کاربران تولید شود. با انجام تحلیل‌های مختلف بر روی این حجم عظیم از داده‌ها می‌توان به اطلاعات مختلفی دست یافت. متون خبری دسته‌ای از این داده‌ها هستند که می‌توان با تحلیل آن‌ها اطلاعات مفیدی را استخراج کرد، از جمله این تحلیل‌ها بررسی میزان ارتباط اخبار با یکدیگر و دسته‌بندی اخبار از جهات گوناگون است. همچنین متون خبری در مقایسه با متون کوتاه شامل کلمات بیشتری هستند که به سبب آن می‌توان از اطلاعات مربوط به هم‌رخدادی کلمات استفاده کرد و همچنین با اجرای روش‌های مدل‌سازی موضوعی بر روی آن‌ها به اطلاعات مفیدی دست یافت. اطلاعات به‌دست آمده از تحلیل متون خبری می‌تواند شامل دانش‌های مفیدی در دنیای واقعی مثل رویدادها و اتفاقات مرتبط در اجتماع باشد.

حوزه‌ی پژوهشی تشخیص موضوع و ردیابی آن (TDT[1]) با هدف انجام این دست از تحلیل‌ها بر روی متون ایجاد گردید و در آن به مسائل مختلفی پرداخته شده است که عموماً هدف آن‌ها بررسی شباهت میان اسناد و ردیابی یک سند است که مشخص می‌کند خبر مربوط به رویداد جدیدی است یا اخبار مرتبط با آن قبلاً هم وجود داشته است [۱]. از جمله مسائل موجود در حوزه‌ی پژوهشی TDT

تشخیص موضوع[۲] است که در این مقاله به آن پرداخته شده و یک روش تشخیص موضوع ارائه می‌شود. در تشخیص موضوع هدف خوشه‌بندی اسناد براساس موضوع آن‌ها از نوع خوشه‌بندی سخت[۳] است به نحوی که هر سند تنها می‌تواند درون یک خوشه قرار گیرد.

پژوهش‌های بسیاری در حوزه‌ی تشخیص موضوع انجام شده است که معمولاً در آن‌ها از روش‌های مختلف خوشه‌بندی استفاده می‌شود و سعی می‌شود با ایجاد تغییراتی از جمله تعیین معیار فاصله میان اسناد، به کار بردن روش‌های جدید مثل استفاده کلید واژه‌ها به جای کل متن و تغییر در ساختار خوشه‌بندی کیفیت تشخیص موضوع را بهبود دهند. همچنین در پژوهش‌هایی سعی بر آن شده تا از روش‌های مدل‌سازی موضوعی مانند LDA[۴] برای تشخیص موضوع استفاده شود.

روش مدل‌سازی موضوعی LDA یک مدل احتمالی تولیدی است و قابل اجرا بر روی دادگان گسسته همانند پیکره‌های متنی است، به نحوی که اسناد را به‌صورت مجموعه‌ای از کلمات در نظر می‌گیرد که ترتیب کلمات در آن اهمیتی ندارد [۲]. در این روش یک مدل بیزی بر روی پیکره‌ای از اسناد ایجاد می‌شود و فرض می‌شود که هر سند می‌تواند موضوعات مختلفی داشته باشد که با توجه به کلمات تشکیل دهنده‌ی هر سند می‌توان سهم هر موضوع را مشخص کرد. از این رو در این روش مقدار معلوم توزیع کلمات در هر سند است و موارد مجهول توزیع موضوع‌ها در هر یک از اسناد و توزیع کلمات در هر یک از موضوعات است که می‌توان این موارد را با استفاده از روش‌هایی مانند روش تغییر[۵] و الگوریتم EM تخمین زد.

در این مقاله روشی برای تشخیص موضوع در متون خبری ارائه می‌شود که بر مبنای الگوریتم خوشه‌بندی K-means است. در این روش برای اجرای خوشه‌بندی K-means ابتدا برای تعیین شباهت میان اسناد خبری از توزیع موضوع حاصل از LDA این اسناد استفاده شده است. با توجه به این‌که الگوریتم K-means به انتخاب مراکز اولیه برای شروع بسیار حساس است [۳]، در این مقاله روشی هوشمندانه بر مبنای مدل‌سازی موضوعی LDA برای انتخاب مراکز اولیه ارائه شده است که تأثیر چشم‌گیری در بهبود کیفیت دارد.

روش کار الگوریتم K-means به این صورت است که ابتدا K سند (k تعداد خوشه‌های نهایی خواهد بود) به عنوان مراکز اولیه به آن داده می‌شود. سپس فاصله سایر اسناد با این مراکز محاسبه شده و هر سند درون نزدیک‌ترین خوشه قرار می‌گیرد. در ادامه زمانی که خوشه‌ها تکمیل شدند سند مرکزی هر خوشه که کمترین مجموع فاصله با سایر اسناد آن خوشه را دارد پیدا می‌شود و پس از به‌روزرسانی مراکز این روال تا زمانی که دیگر هیچ خوشه‌ای تغییر نکند ادامه می‌یابد.

در این مقاله به بررسی میزان حساسیت خوشه‌بندی K-means به انتخاب مراکز اولیه در مسئله تشخیص موضوع پرداختیم و با آزمایش بر روی مجموعه دادگان مختلف نشان دادیم که انتخاب مراکز اولیه در کارایی این الگوریتم تأثیر بسیاری دارد. از این رو روشی بر مبنای مدل‌سازی LDA ارائه کردیم که به انتخاب هوشمندانه‌ی مراکز اولیه می‌پردازد و باعث بهبود کیفیت می‌شود. همچنین با توجه به بی ناظر بودن مسئله برای تعیین پارامتر مربوط به روش پیشنهادی روشی ارائه کرده‌ایم که با بهره‌گیری از یک معیار بی ناظر برای تعیین کیفیت در خوشه‌بندی به نام Silhouette و بدون نیاز به دادگان طلایی می‌تواند پارامتر مناسب را پیدا کند و روش پیشنهادی در مجموع بی ناظر باقی بماند.

ازجمله ویژگی‌های اصلی و دست‌آوردهای روش ارائه شده می‌توان به موارد زیر اشاره کرد:

- بررسی رفتار الگوریتم K-means در مسئله تشخیص موضوع و نقش تعیین مراکز اولیه در میزان کارایی این الگوریتم
- استفاده از توزیع موضوع حاصل از LDA اسناد برای محاسبه شباهت میان اسناد و ارائه روش جدید برای پیدا کردن مراکز اولیه الگوریتم K-means بر مبنای توزیع موضوع حاصل از LDA اسناد
- تنظیم مقدار پارامتر در الگوریتم انتخاب مراکز اولیه بدون نیاز به دادگان طلایی و بی ناظر بودن کل روش پیشنهادی
- بهبود چشم‌گیر کیفیت تشخیص موضوع نسبت به روش LDA در دو مجموعه از سه مجموعه دادگان مورد آزمایش با انتخاب مراکز اولیه‌ی مناسب‌تر

در ادامه در بخش دوم به مروری از پژوهش‌های انجام شده در حوزه‌ی تشخیص موضوع و همچنین روش‌های ارائه شده برای انتخاب مراکز اولیه در الگوریتم K-means پرداخته می‌شود. در بخش سه مراحل روش پیشنهادی ارائه می‌شود و جزئیات هر یک از مراحل را بیان می‌کنیم. در بخش چهار ابتدا سه مجموعه دادگان استاندارد در مسئله تشخیص موضوع را معرفی می‌کنیم همچنین معیارهای ارزیابی که از آن‌ها استفاده شده است را معرفی می‌کنیم، در ادامه‌ی این بخش به بیان جزئیات آزمایش‌ها و نتایج به‌دست آمده می‌پردازیم و آن‌ها را با نتایج حاصل از روش‌های رقیب مقایسه می‌کنیم و در نهایت به تحلیل نتایج می‌پردازیم. نهایتاً در بخش پنجم به جمع‌بندی و نتیجه‌گیری مقاله پرداخته شده است.

۲- مروری بر کارهای پیشین

در حوزه تشخیص موضوع پژوهش‌های گوناگونی انجام شده است و روش‌های مختلفی ارائه شده است. دسته‌ای از روش‌ها با استفاده از الگوریتم‌های خوشه‌بندی و با ایجاد تغییراتی برای بهبود آن‌ها اقدام به خوشه‌بندی اسناد می‌کنند. در برخی از روش‌ها برای محاسبه فاصله میان اسناد تمامی محتویات سند در نظر گرفته می‌شود. پژوهش [۴] از جمله این روش‌ها است که برای خوشه‌بندی فاصله کسینوسی میان اسناد را محاسبه می‌کند؛ اما در برخی از روش‌ها با توجه به هزینه محاسباتی زیاد برای در نظر گرفتن کل متن سند، ابتدا کلمات کلیدی را استخراج و سپس اقدام به خوشه‌بندی می‌کنند. برای نمونه پژوهش [۵] پس از استخراج کلمات کلیدی، با استفاده از تحلیل Wavelet اقدام به خوشه‌بندی می‌کند. همچنین در مقاله [۶] از تشخیص موضوع برای پیش‌بینی میزان محبوبیت یک خبر استفاده شده و برای تشخیص موضوع از خوشه‌بندی مبتنی بر کلمات کلیدی استفاده شده است.

در پژوهش [۷] روشی ارائه شده است که کاربرد آن در دسته‌بندی اخبار مرتبط به هم در کنار ویدیوهای مربوط به این اخبار است. در این روش در قسمت مربوط به تشخیص موضوع گراف هم رخدادی کلمات اسناد را ساخته و شباهت اسناد را براساس آن محاسبه می‌کند سپس با داشتن میزان شباهت اقدام به خوشه‌بندی می‌کند. در این روش برای پیدا کردن ویدیوهای مرتبط نیز شباهت میان تصویر خبرها را با فریم اصلی ویدیوها بررسی می‌کند. در پژوهش [۸] به کاربرد دیگر تشخیص موضوع پرداخته شده است که در آن هدف این است که مطالب منتشر شده دسته‌بندی شوند و موضوعات آن‌ها مشخص شود، سپس بررسی می‌شود که آیا این مطالب مرتبط با شرکت خاصی است یا نه. با دادن این اطلاعات به کارشناسان شرکت‌ها آن‌ها می‌توانند در رابطه با تأثیر مطالب منتشر شده بر روی اعتبار شرکت‌ها تصمیم‌گیری کنند. روش این مقاله به این صورت است که ابتدا با استفاده از برچسب‌های موجود داده‌ها را رده‌بندی می‌کند و از وزن‌های به‌دست آمده برای خوشه‌بندی داده‌ها استفاده می‌کند.

دسته‌ی دیگری از پژوهش‌ها در حوزه‌ی تشخیص موضوع از روش‌های مدل‌سازی موضوعی و به ویژه LDA استفاده می‌کنند. برای نمونه در پژوهش [۹] به این ویژگی که هر رویداد و موضوع در بازه‌ی زمانی خاصی اتفاق می‌افتد توجه

شده و با استفاده از LDA مدلی ارائه شده است که در آن زمان نیز در نظر گرفته می‌شود. در پژوهش [۱۰] علاوه بر مدل LDA مدل AT[۶] [۱۱] را بررسی می‌کند که در این مدل علاوه بر توزیع‌های موجود در مدل LDA توزیع موضوعات مختلف برای هر کاربر محاسبه می‌شود، سپس این پژوهش مدلی را ارائه می‌دهد که با توجه به توزیع کلمات برای هر کاربر عمل می‌کند. در مقاله [۱۲] هدف این است که در شبکه‌های اجتماعی به جای استفاده از زمان و ساختار شبکه‌ی اجتماعی برای نمایش مطالب، ابتدا آن‌ها را دسته‌بندی کرده و سپس نمایش دهیم. برای تشخیص موضوع در این پژوهش بر مبنای مدل Labeled LDA [۱۳] عمل شده و از برچسب‌های موجود در مجموعه دادگان استفاده شده است.

پژوهش [۱۴] از مدل‌سازی موضوعی برای تشخیص موضوع در متون مکالمه‌ای و محاوره‌ای استفاده کرده است. روش ارائه شده در این پژوهش یک مدل گسترش یافته LDA است که به جای استفاده از کلمات به سطح جمله آمده است. در این روش برای جلوگیری از تنکی داده‌ها از روابط جزبه کل[۷] استفاده کرده و در نهایت مدل LDA را به‌گونه‌ای تغییر داده که شامل ویژگی‌های گفتاری و روابط جز به کل شود. در مقاله [۱۵] یک روش با نظارت برای دسته‌بندی اسناد با استفاده از LDA ارائه شده است که در آن برای پیدا کردن کلمات هم‌معنی و با کاربرد مشابه از مدل‌سازی موضوعی LDA استفاده کرده است.

همان‌طور که گفته شد یکی از قسمت‌های مهم این مقاله ارائه‌ی روش جدیدی برای تعیین مراکز اولیه در الگوریتم K-means است، در رابطه با مسئله تعیین مراکز اولیه‌ی مناسب برای الگوریتم K-means نیز پژوهش‌های گوناگونی انجام شده و روش‌های مختلفی ارائه شده است. در روش ارائه شده در مقاله [۱۶] پیشنهاد شده است که K نمونه را از داده‌ها به‌صورت تصادفی به عنوان مراکز اولیه انتخاب شود که ایده اصلی از این کار این است که زمانی که نقاط به‌صورت تصادفی انتخاب می‌شوند به احتمال زیاد این نقاط انتخاب شده در قسمت‌های پرتراکم قرار داشته‌اند. در روشی که در مقاله [۱۷] ارائه شده است ابتدا یکی از داده‌ها به‌صورت تصادفی به عنوان اولین مرکز انتخاب می‌شود. سپس برای انتخاب مراکز بعدی فاصله هر یک از داده‌ها با نزدیک‌ترین نقطه از میان مراکزی که قبلاً انتخاب شده‌اند محاسبه می‌شود و نمونه‌ای که بیشترین فاصله را دارد به مجموعه‌ی مراکز انتخاب شده اضافه می‌شود، این روند تا انتخاب K مرکز ادامه پیدا می‌کند.

یکی از روش‌های معروف برای انتخاب مراکز اولیه روش K-means++ [۱۸] است، در این روش هدف این است که احتمال انتخاب مراکز متناسب با میزان فاصله آن‌ها با مراکزی که قبلاً انتخاب شده‌اند باشد. اگر مجموعه دادگان را $X = \{x_1, x_2, \dots, x_N\}$ در نظر بگیریم که شامل N داده است، در این روش مرکز اول را به‌صورت تصادفی انتخاب می‌کنیم. برای انتخاب اُمین مرکز که $i \in \{2,3, \dots, K\}$ است $x' \in X$ با احتمال $\frac{md(x')^2}{\sum_{j=1}^{N} md(x_j)^2}$ انتخاب می‌شود، که در آن $md(x)$ برابر با فاصله داده x با نزدیک‌ترین مرکز از میان مراکزی که قبلاً انتخاب شده‌اند است. در این روش همانند روش پیشنهادی ما هدف انتخاب مراکز اولیه مناسب برای الگوریتم K-means است که موجب بهبود کیفیت خوشه‌بندی می‌شود و استفاده از این روش برای تعیین مراکز اولیه بسیار رایج است، از این رو در این مقاله کیفیت روش پیشنهادی را با این روش نیز مقایسه کرده‌ایم.

در روش ارائه شده در [۱۹] ابتدا K مؤلفه‌ی مستقل[۸] از مجموعه دادگان را محاسبه کره و سپس K مرکز را به نحوی انتخاب می‌کند که این مراکز کمترین فاصله‌ی کسینوسی را با مؤلفه‌های مستقل داشته باشند. در مقاله [۲۰] ابتدا مرکز کل داده‌ها را محاسبه می‌کند و نقاط را به ترتیب فاصله‌ی آن‌ها تا مرکز مرتب می‌کند. برای انتخاب اُمین مرکز که $i \in \{1,2,3, \dots, K\}$ است نقطه‌ی شماره‌ی $1 + (i-1)N/K$ را از ترتیب جدید انتخاب می‌کند، این امر به پراکندگی داده‌های انتخاب شده کمک می‌کند. در مقاله [۲۱] نیز پیشنهاد شده است که ابتدا داده‌ها را با استفاده از روش سلسله مراتبی خوشه‌بندی کرده سپس مراکز به‌دست آمده از خوشه‌های این روش را به عنوان مراکز اولیه الگوریتم K-means در نظر گرفته و اقدام به خوشه‌بندی کنیم.

۳- روش پیشنهادی

در این بخش مراحل مختلف روش پیشنهادی و جزئیات قسمت‌های مختلف آن توضیح داده می‌شود. در روش پیشنهادی ابتدا توزیع موضوع هر یک از اسناد مجموعه دادگان توسط LDA محاسبه می‌شود و از این توزیع به عنوان بردار نماینده‌ی اسناد استفاده می‌شود. در مرحله‌ی بعد، با استفاده از این بردارها، الگوریتمی برای پیدا کردن مراکز اولیه مناسب برای روش خوشه‌بندی K-means ارائه می‌شود. در نهایت الگوریتم خوشه‌بندی K-means بر روی بردارهای نماینده‌ی اسناد و با در نظر گرفتن مراکز اولیه‌ی به‌دست آمده از مرحله‌ی قبل اجرا می‌شود. الگوریتم ارائه شده برای پیدا کردن مراکز اولیه مناسب به یک پارامتر آستانه‌ای وابسته است؛ بنابراین تنظیم این پارامتر به‌منظور دستیابی به کیفیت بالا ضروری است. با توجه به بی ناظر بودن مسئله‌ی تشخیص موضوع، برای تعیین مقدار این پارامتر نمی‌توان از داده‌های برچسب خورده استفاده کرد. از این رو در مرحله‌ی نهایی روشی را ارائه کرده‌ایم که با استفاده از آن و بدون نیاز به داده‌های برچسب خورده مقدار مناسب برای پارامتر آستانه معین شود. در ادامه‌ی این بخش جزئیات مراحل روش را بیان می‌کنیم.

۳-۱- محاسبه توزیع موضوع حاصل از LDA و فاصله میان اسناد

همان‌طور که گفته شد الگوریتم K-means براساس فاصله میان نقاط در فضا عمل می‌کند و اسناد را درون خوشه‌ای قرار می‌دهد که با مرکز آن خوشه کمترین فاصله را داشته باشند، سپس در هر خوشه سندی که کمترین فاصله با اسناد دیگر دارد را به عنوان مرکز جدید انتخاب می‌کند. پس از تعیین مراکز جدید مجدداً هر سند درون نزدیک‌ترین خوشه قرار می‌گیرد و این عمل تا زمانی که دیگر هیچ خوشه‌ای تغییر نکند تکرار می‌شود. از این رو برای تعیین فاصله نیاز به برداری از ویژگی‌ها برای هر سند است که بتوان براساس آن فاصله اسناد را تعیین کرد. در روش‌های پیشین معمولاً برای تعیین فاصله از بردار کلمات استفاده می‌شد اما در این مقاله با توجه به ویژگی‌های مفید LDA در مسئله تشخیص موضوع ابتدا توزیع LDA را که نشان‌دهنده‌ی توزیع موضوعات در یک سند است محاسبه کرده و برای تعیین فاصله میان اسناد فاصله میان توزیع‌های آن‌ها را محاسبه می‌کنیم.

پس از محاسبه توزیع موضوع حاصل از LDA هر سند نیاز است تا انتخاب با یکی از معیارهای فاصله و اعمال آن‌ها بر روی توزیع‌ها فاصله اسناد محاسبه شود. از این رو آزمایشات را با معیارهای مختلفی از جمله فاصله اقلیدسی و کسینوسی انجام دادیم که نتایج نشان داد استفاده از معیار شباهت کسینوسی برای محاسبه میزان فاصله میان اسناد مناسب‌تر بوده و نتایج بهتری را به همراه دارد. همچنین با توجه به این نکته که در الگوریتم ارائه شده در این مقاله برای پیدا کردن مراکز اولیه مناسب بخش ۳-۲ روش پیشنهادی براساس توزیع LDA عمل می‌کند از توزیع‌های محاسبه شده در این مرحله استفاده می‌شود.

۳-۲- پیدا کردن مراکز اولیه‌ی مناسب و خوشه‌بندی

کیفیت روش خوشه‌بندی K-means به انتخاب مراکز اولیه بسیار حساس است [۳]. همچنین آزمایش‌هایی که ما در مسئله تشخیص موضوع انجام داده‌ایم نشان

می‌دهند این الگوریتم حساسیت نسبتاً بالایی به مراکز اولیه در این مسئله نیز دارد بخش ۴-۳. از این رو روش جدیدی را ارائه کردیم که بر مبنای توزیع LDA اسناد عمل کرده و مراکز مناسب را برای خوشه‌بندی K-means پیدا می‌کند.

پس از محاسبه توزیع LDA که به هر سند یک بردار K (تعداد موضوع‌ها) بعدی اختصاص می‌دهد، K مجموعه خالی $s_1, s_2, s_3, \dots, s_K$ ایجاد می‌کنیم. برای هر یک از بردارهای متناظر با اسناد، مقدار max/max2 را محاسبه می‌کنیم که در آن max مقدار بُعد بیشینه و max2 برابر با مقدار دومین بیشینه در میان ابعاد است. اگر در سندی بعد iام بیشینه باشد و دومین مقدار بیشینه مربوط به بعد jام باشد، با در نظر گرفتن مقدار آستانه t (که مقدار مناسب آن به‌صورت بی ناظر با روشی که در بخش ۳-۳ توضیح می‌دهیم پیدا می‌شود) بررسی می‌کنیم که آیا max/max2 بیشتر از مقدار آستانه t است یا خیر. در صورتی‌که بیشتر بود آن سند را به مجموعه‌ی s_i (i اندیس بُعد بیشینه در سند مورد نظر است) اضافه می‌کنیم. این روند را برای تمامی اسناد تکرار می‌کنیم تا تمامی مجموعه‌ها تکمیل شوند.

در مرحله بعد مرکز هر یک از مجموعه‌های $s_1, s_2, s_3, \dots, s_K$ را پیدا می‌کنیم که برای این کار کافی است میانگین توزیع‌های موضوع حاصل از LDA را برای اسناد موجود در هر سند محاسبه کنیم. سپس مراکز به‌دست آمده را به عنوان مراکز اولیه به الگوریتم K-means برای خوشه‌بندی می‌دهیم. این الگوریتم نیز هر سند را درون خوشه‌ای قرار می‌دهد که کمترین فاصله را تا مرکز آن دارند. برای محاسبه فاصله در اینجا از فاصله کسینوسی میان توزیع موضوع حاصل از LDA اسناد استفاده شده است. پس از تکمیل خوشه‌ها مراکز جدید خوشه محاسبه می‌شوند مجدداً خوشه‌ی مربوط به هر سند براساس مراکز جدید شناسایی می‌شود. این روند تا زمانی که دیگر هیچ خوشه‌ای تغییر نکند ادامه پیدا می‌کند.

در این روش ابتدا مقدار K برابر با تعداد موضوعات قرار داده می‌شود و به همین تعداد مجموعه خالی ایجاد می‌کنیم. پس از انجام روند افزودن نمونه‌ها به این مجموعه‌ها با توجه به انتخاب مقدار پارامتر آستانه برخی از مجموعه‌ها خالی باقی می‌مانند. با توجه تحلیل‌های انجام شده بر روی مجموعه‌های دادگان بخش ۴-۴ توزیع اسناد در موضوعات مختلف در دو مجموعه از سه مجموعه دادگان مورد ارزیابی توانی است و تعداد زیادی از اسناد در چند خوشه قرار دارند و بسیاری از خوشه‌ها دارای تعداد کمی سند هستند. از این رو در روش ارائه شده تعداد خوشه‌ها کاهش می‌یابد به نحوی که خوشه‌های مطمئن باقی می‌مانند و خوشه‌هایی که اهمیت پایینی دارند حذف می‌شوند. همان‌طور که مشخص است هرچه مقدار پارامتر آستانه را افزایش دهیم تعداد خوشه‌های نهایی کاهش می‌یابد.

از نظر پیچیدگی زمانی روش ارائه شده این‌گونه است که ابتدا یک بار کل داده‌ها پیمایش می‌شوند تا اسنادی که درون مجموعه‌های $s_1, s_2, s_3, \dots, s_K$ قرار می‌گیرند شناسایی شوند. پس از آن یک بار عمل میانگین‌گیری بر روی داده‌های بازمانده در مجموعه‌ها (که در بدترین حالت برابر با تعداد کل داده‌ها است) انجام می‌شود. با توجه به اینکه پیچیدگی زمانی هر دو مرحله خطی است، روش ارائه شده برای پیدا کردن مراکز اولیه مناسب در کل پیچیدگی زمانی خطی دارد.

۳-۳- تعیین مقدار آستانه مناسب به‌صورت بی ناظر

با توجه به این نکته که تشخیص موضوع یک مسئله‌ی بی ناظر است برای تعیین مقدار مناسب پارامتر آستانه t که در قسمت قبل توضیح داده شد نمی‌توان از داده‌های برچسب خورده موجود در مجموعه دادگان طلایی استفاده کرد. از این رو نیاز به یک معیار بی ناظر است که بتواند با توجه به ویژگی‌های درونی داده‌ها عمل کرده و ما را در تنظیم مقدار پارامتر یاری کند.

معیارهای مختلفی وجود دارند که برای ارزیابی مسئله‌ی خوشه‌بندی از آن‌ها استفاده می‌شود و می‌توانند به‌صورت بی ناظر و بدون نیاز به دادگان برچسب خورده کیفیت خوشه‌بندی را بررسی کنند. از این رو برای تعیین مقدار پارامتر آستانه به بررسی نتایج حاصل برای سه معیار Beta-CV، برش نرمال و Silhouette [۲۲] پرداختیم. معیار Beta-CV به‌صورت تقسیم میانگین فاصله درون خوشه‌ای (فاصله هر نمونه با نمونه‌های دیگر همان خوشه‌ای که در آن قرار دارد) بر میانگین فاصله بین خوشه‌ها (فاصله نمونه با نمونه‌های خوشه‌های دیگر) محاسبه می‌شود. طبیعی است که هرچه این مقدار کمتر باشد خوشه‌بندی کیفیت بهتری دارد. معیار برش نرمال برابر است با مجموع نسبت فاصله بین خوشه‌ای به حجم خوشه به ازای هر خوشه. حجم خوشه نیز به‌صورت مجموع فاصله‌های نمونه‌های درون خوشه با تمام نمونه‌ها محاسبه می‌شود. این معیار وزن برش‌ها را محاسبه می‌کند و هرچه مقدار آن بیشتر باشد نشان‌دهنده‌ی خوشه‌بندی بهتر است. معیار سوم معیار Silhouette است که در ادامه به طور مفصل توضیح داده خواهد شد.

آزمایش‌های ما نشان می‌دهند که معیار Silhouette عملکرد بهتری نسبت به دو معیار دیگر دارد از این‌رو در روش پیشنهادی ما برای پیدا کردن مقدار مناسب برای پارامتر آستانه از معیار Silhouette استفاده شده است. با توجه به تحلیل‌هایی که بر روی مجموعه‌های دادگان انجام دادیم بخش ۴-۴ توزیع اسناد در دو مجموعه دادگان از سه مجموعه دادگان مورد آزمایش به‌صورت توانی است و معیار Silhouette در مجموعه‌های دادگان با توزیع توانی می‌تواند عملکرد بهتری نسبت به دو معیار دیگر داشته باشد. در ادامه به توضیح منطق معیار Silhouette و روش محاسبه‌ی آن می‌پردازیم.

هدف معیار Silhouette [۲۳] کمتر بودن فاصله میان اعضای یک خوشه و زیاد بودن فاصله هر نمونه با خوشه‌های دیگر است. در این معیار مقدار a(i) میانگین فاصله عضو iام با سایر اعضای خوشه‌ای که در آن حضور دارد و مقدار b(i) نیز میانگین فاصله عنصر iام با عناصر متعلق به خوشه‌ی نزدیک‌ترین همسایه عنصر iام تعریف می‌شوند. برای پیدا کردن خوشه همسایه عنصر iام میانگین فاصله این عنصر با تک تک خوشه‌ها به غیر از خوشه‌ای که در آن قرار دارد محاسبه می‌شود و نزدیک‌ترین خوشه به عنوان خوشه همسایه انتخاب می‌شود. مقدار Silhouette برای عنصر iام به صورتی که در فرمول شماره ۱ آمده است محاسبه می‌شود. با توجه به این فرمول مشخص است که مقدار S(i) همواره عددی بین ۱ و ۱- است، این مقدار در صورتی به ۱ نزدیک می‌شود که مقدار a(i) بسیار کوچک‌تر از b(i) باشد. کم بودن مقدار a(i) نشان می‌دهد که فاصله عنصر iام با عناصر خوشه‌ی خود بسیار کم است و زیاد بودن مقدار b(i) نشان می‌دهد که فاصله عنصر iام با عناصر خوشه‌ی همسایه زیاد است، در نتیجه نزدیک بودن مقدار S(i) به ۱ نشان دهنده‌ی این است که این عنصر در خوشه‌ی صحیحی قرار گرفته است. به همین ترتیب می‌توان نتیجه گرفت که اگر مقدار S(i) به ۱- نزدیک شود عنصر iام در خوشه مناسبی قرار ندارد و بهتر بود این عنصر در خوشه‌ی همسایه قرار گیرد.

$$S(i) = \frac{b(i)-a(i)}{Max\{b(i),a(i)\}} \qquad (۱)$$

برای محاسبه ضریب Silhouette کل، میانگین را با استفاده از فرمول ۲ محاسبه می‌کنیم. مقدار Silhouette در بهترین حالت برابر ۱ و در بدترین حالت برابر با ۱- است.

$$SC = \frac{1}{n}\sum_{i=1}^{n} S(i) \qquad (۲)$$

در روش پیشنهادی ما هر چه مقدار پارامتر آستانه را افزایش دهیم تعداد خوشه‌ها کاهش پیدا می‌کنند. با توجه به این نکته که در معیار Silhouette و سایر معیارهای خوشه‌بندی که در آن‌ها به فاصله نمونه‌ها درون یک خوشه اهمیت داده می‌شود، هر چه تعداد خوشه‌ها افزایش یابد این معیارها نیز بهبود می‌یابند. بنابراین افزایش این معیارها می‌تواند به دو دلیل انجام شود: دلیل اول زیاد شدن تعداد خوشه‌ها و دلیل دوم بهبود در خوشه‌بندی. از این رو منطقی است در استفاده از این معیارها، به جای در نظر گرفتن مقدار بیشینه، نقطه‌ای از نمودار که مقدار معیار در آن دچار تغییر و شکستی شده است در نظر گرفته شود. زیرا این نقطه می‌تواند محلی باشد که در آن بهبود تنها به دلیل زیاد بودن خوشه‌ها نبوده و کیفیت خوشه‌بندی در آن بهبود یافته است.

از این رو با توجه به این‌که تعداد خوشه‌ها بر روند تغییر Silhouette تأثیر می‌گذارد و هرچه تعداد خوشه‌ها بیشتر باشد مقدار Silhouette نیز بیشتر خواهد بود، برای پیدا کردن مقدار پارامتر مناسب نمی‌توان صرفاً پارامتر متناظر با بیشینه‌ی Silhouette را انتخاب کرد و نیاز به بررسی روند تغییرات نمودار Silhouette و تصمیم‌گیری براساس آن داریم. در روش پیشنهادی برای تعیین مقدار آستانه نقطه‌ای از نمودار را که در آنجا با بیشترین میزان شکستگی و کاهش Silhouette روبه‌رو هستیم انتخاب کرده و مقدار پارامتر آستانه را از روی آن تشخیص می‌دهیم. در ادامه آزمایش‌های انجام شده و نتایج حاصل از آن ارائه خواهند شد.

۴- آزمایش‌ها و ارزیابی نتایج

۴-۱- مجموعه دادگان

برای ارزیابی و تحلیل نتایج به‌دست آمده، آزمایش‌های مربوط به روش پیشنهادی و رقیب‌هایی که قرار است با روش پیشنهادی مقایسه شوند را بر روی ۳ مجموعه دادگان معتبر و استاندارد در مسئله تشخیص موضوع یعنی TDT2، Reuters-21578 و 20 newsgroups انجام دادیم [۲۷-۲۴]. مجموعه دادگان TDT2 برگرفته از اخبار در موضوعات مختلف است که در ۹۶ دسته موضوع دسته‌بندی شده است [۲۸].

اسناد خبری موجود در این مجموعه دادگان می‌توانند دارای یک و یا بیش از یک موضوع باشند و با توجه به این‌که در مسئله تشخیص موضوع هر سند تنها می‌توان متعلق به یک موضوع باشد همانند پژوهش‌های مشابهی که قبلاً انجام شده‌اند مانند [۲۷] و [۲۹] اسناد دارای بیش از یک موضوع را حذف کرده و همچنین اسناد مربوط به ۳۰ دسته‌ای که بیشترین تعداد خبر را دارند در نظر گرفته‌ایم.

مجموعه دادگان دوم Reuters-21578 است که این مجموعه دادگان شامل ۱۳۵ دسته موضوع می‌باشد که در این مورد نیز اسناد دارای چند موضوع را حذف کرده و تعداد موضوعات از ۱۳۵ به ۶۵ موضوع کاهش یافته است [۳۰]. مجموعه دادگان سوم که مورد استفاده قرار دادیم 20 newsgroups [۳۱] است که از نسخه‌ی مرتب شده آن استفاده کردیم، در این نسخه قسمت‌های اضافی مانند مشخصات گروه خبری حذف شده است. در این مجموعه دادگان اسناد در ۲۰ دسته خبری دسته‌بندی شده‌اند و هر خبر تنها یک موضوع دارد. با توجه به این نکته که مسئله از نوع خوشه‌بندی سخت است و در مجموعه دادگان مختلف باید اسناد دارای بیش از یک موضوع را حذف کنیم و این امر موجب کاهش تعداد اسناد و موضوعات می‌شود، اطلاعات مربوط به مجموعه دادگان‌های استفاده شده قبل و بعد از تغییرات را در جدول شماره ۱ مشاهده می‌کنید.

۴-۲- معیارهای ارزیابی

برای ارزیابی و مقایسه روش ارائه شده از سه معیار دقت، فراخوانی و معیار F استفاده می‌کنیم. نحوه محاسبه این معیارها در مسئله تشخیص موضوع را در ادامه به نحوی که در [۳۲] نیز بیان شده است توضیح می‌دهیم. مجموعه داده‌های صحیح و برچسب خورده را $C^* = \{c_1^*, c_2^*, c_3^*, \dots, c_k^*\}$ در نظر می‌گیریم که در آن K تعداد کل خوشه‌ها و c_i^* نشان دهنده‌ی خوشه iام در مجموعه دادگان طلایی است. به همین ترتیب مجموعه $C = \{c_1, c_2, c_3, \dots, c_k\}$ را به عنوان مجموعه خوشه‌های حاصل از روش پیشنهادی در نظر می‌گیریم که در آن c_i خوشه‌ی iام ایجاد شده از روش پیشنهادی است. حال می‌توان برای هر جفت خوشه از مجموعه C و C^* این سه معیار را محاسبه کرد، به نحوی که حاصل تقسیم میزان اشتراک خوشه حاصل از مجموعه دادگان طلایی و روش مورد ارزیابی بر تعداد اعضای خوشه مربوط به روش مورد ارزیابی نشان دهنده‌ی معیار دقت و حاصل تقسیم این میزان اشتراک بر تعداد اعضای خوشه‌ی مربوط به دادگان طلایی نشان دهنده‌ی فراخوانی است. همچنین با داشتن این دو معیار می‌توان معیار F را نیز محاسبه کرد که در زیر فرمول‌های مربوط به محاسبه‌ی این معیارها آمده است.

جدول ۱- اطلاعات مجموعه‌های دادگان

مجموعه دادگان	20 newsgroups	Reuters 21578	TDT2
تعداد کل اسناد	18846	21578	11201
تعداد کل موضوعات	20	135	96
تعداد اسناد پس از حذف اسناد چند موضوعی	18846	8293	9394
تعداد موضوعات پس از حذف اسناد چند موضوعی	20	65	30

$$\text{precision}(c_j, c_i^*) = |c_j \cap c_i^*| / |c_j| \quad (۳)$$

$$\text{recall}(c_j, c_i^*) = |c_j \cap c_i^*| / |c_i^*| \quad (۴)$$

$$F(c_j, c_i^*) = \frac{2\text{precision}(c_j, c_i^*).\text{recall}(c_j, c_i^*)}{\text{precision}(c_j, c_i^*)+\text{recall}(c_j, c_i^*)} \quad (۵)$$

هر کدام از خوشه‌های موجود در مجموعه C^* را در کنار تک تک خوشه‌های مجموعه C قرار می‌دهیم و مقدار معیار F را به صورتی که گفته شد محاسبه می‌کنیم، خوشه‌ای از مجموعه C را که بیشترین مقدار معیار F را دارد را به عنوان خوشه نظیر به آن خوشه در مجموعه طلایی در نظر می‌گیریم. در نهایت میانگین معیارهای دقت، فراخوانی و معیار F را با استفاده از فرمول‌های زیر محاسبه می‌کنیم:

$$\text{precision} = \sum_{c^* \in C^*} \frac{|c^*|}{\sum_{c^* \in C^*} |c^*|} P(c^*) \quad (۶)$$

$$\text{Recall} = \sum_{c^* \in C^*} \frac{|c^*|}{\sum_{c^* \in C^*} |c^*|} R(c^*) \quad (۷)$$

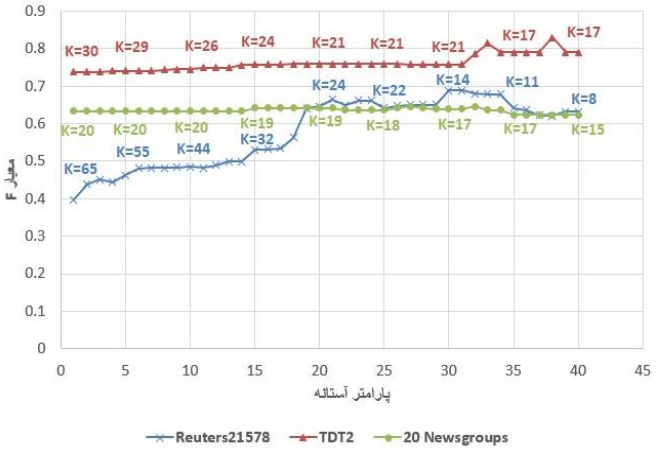
$$F - \text{value} = \frac{2 \text{ precision} . \text{ recall}}{\text{precision}+\text{recall}} \quad (۸)$$

۴-۳- نتایج

در این بخش جزئیات روند انجام آزمایش‌ها را توضیح داده و به بیان نتایج و مقایسه آن‌ها می‌پردازیم. همان‌طور که گفته شد در ابتدا نیاز به محاسبه توزیع موضوع حاصل از LDA است که به این منظور با بهره‌گیری از [۳۳] و [۳۴] الگوریتم LDA را بر روی دادگان اجرا کردیم. خروجی این مرحله یک ماتریس n×k برای هر یک از مجموعه‌های دادگان است که n تعداد اسناد آن مجموعه دادگان و k تعداد موضوعات است. برای مقداردهی پارامترهای اولیه الگوریتم LDA یعنی α و β نیز از مقادیر رایج آن‌ها استفاده کردیم و پارامتر α آن را برابر 0.5 و β را برابر با 0.1 قرار دادیم.

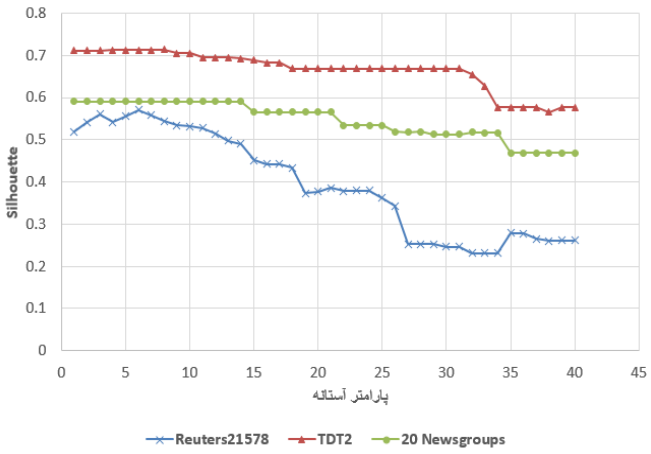
در مرحله بعد الگوریتم پیشنهادی برای پیدا کردن مراکز اولیه‌ی مناسب بخش ۳-۲ را با داشتن توزیع موضوع حاصل از LDA اسناد به ازای پارامترهای آستانه مختلف (از ۱ تا ۴۰) اجرا کردیم. با توجه به این‌که با افزایش پارامتر آستانه تعداد خوشه‌ها کاهش می‌یابد و کاهش بیش از حد تعداد خوشه‌ها باعث بی معنی شدن نتایج می‌شود، میزان پارامتر آستانه را از ۴۰ بالاتر نبرده و نتایج را تا این مقدار بررسی می‌کنیم.

سپس با داشتن مراکز اولیه برای پارامترهای آستانه مختلف و با محاسبه فاصله کسینوسی میان توزیع موضوع حاصل از LDA اسناد اقدام به خوشه‌بندی به روش K-means می‌کنیم. پس از انجام خوشه‌بندی می‌توان مقدار معیار F را برای نتایج مختلف حاصل از مقادیر پارامتر آستانه مختلف محاسبه کرد. نمودار مربوط به تغییرات معیار F را برای هر سه مجموعه دادگان در شکل ۱ ملاحظه می‌کنید، در این نمودار محور افقی مقدار پارامتر آستانه و محور عمودی میزان معیار F می‌باشد، همچنین عدد K (تعداد خوشه‌ها) در مقادیر مختلف آستانه در بالای نمودار نوشته شده است.



شکل ۱- معیار F برای پارامتر آستانه‌های مختلف

با توجه به نتایج حاصل از اجرای روش‌های رقیب بر روی مجموعه‌های دادگان که در ادامه‌ی این بخش آمده است مقادیر معیار F به‌دست آمده برای پارامترهای آستانه مختلف در اکثر نقاط از مقدار معیار F رقیب‌ها بیشتر بوده و تنها در برخی نقاط ممکن است کمتر باشد از طرفی نیاز است که مقدار مشخصی را برای پارامتر آستانه پیدا کرده و نتایج حاصل از آن را با سایر رقیب‌ها مقایسه کرد. به همین منظور با استفاده از روشی که در بخش ۳-۳ توضیح داده شد، مقدار Silhouette را برای مقادیر مختلف پارامتر آستانه محاسبه و نقطه‌ای که در آن بیشترین شکستگی و کاهش را داریم انتخاب می‌کنیم. در شکل ۲ نمودار تغییرات مقدار Silhouette را برای مقادیر مختلف پارامتر آستانه ملاحظه می‌کنید که برای هر ۳ مجموعه دادگان رسم شده است.



شکل ۲- Silhouette برای پارامتر آستانه‌های مختلف

همان‌طور که در شکل ۲ مشخص است مقدار مناسب پارامتر آستانه برای مجموعه دادگان Reuters21578 برابر با عدد ۲۶، برای مجموعه دادگان TDT2 برابر با ۳۳ و برای مجموعه دادگان Newsgroups 20 برابر ۳۴ است. با داشتن این مقادیر می‌توان مقادیر معیارهای ارزیابی را در این مقادیر آستانه با سایر روش‌ها مقایسه کرد. اولین روشی که نتایج آن را با نتایج روش پیشنهادی مقایسه می‌کنیم روش LDA است که بنابر دانش ما بهترین کارایی را در مسئله تشخیص موضوع دارد. در این روش پس از محاسبه توزیع موضوع حاصل از LDA انتخاب خوشه‌ی هر سند براساس اعداد توزیع موضوع حاصل از LDA آن سند صورت می‌گیرد و سند به خوشه‌ای می‌رود که در توزیع موضوع حاصل از LDA آن سند بیشترین مقدار را دارد. مقایسه نتایج این روش با روش پیشنهادی در جدول شماره ۲ آمده است، همان‌طور که مشخص است معیار F روش پیشنهادی نسبت به روش LDA در مجموعه دادگان Reuters21578 و TDT2 به ترتیب ۳۹ و ۹ درصد افزایش داشته و تنها در مجموعه دادگان 20 newsgroups به اندازه‌ی ۰٫۵ درصد کاهش داشته که دلیل کاهش در این مجموعه دادگان خاص را در بخش تحلیل بخش ۴-۴ بررسی می‌کنیم.

جدول ۲- مقایسه نتایج به دست آمده با روش LDA

		Reuters 21578	TDT2	20 newsgroups
روش LDA	تعداد خوشه‌ها	65	30	20
	معیار F	0.463	0.741	0.639
روش پیشنهادی	تعداد خوشه‌ها	19	19	17
	معیار F	0.647	0.814	0.635
درصد میزان افزایش نسبی	درصد میزان افزایش نسبی	39.62%	09.75%	-00.56%

در ادامه نتایج روش پیشنهادی را با دو روش انتخاب مراکز اولیه برای الگوریتم K-means یعنی انتخاب مراکز اولیه به‌صورت تصادفی و روش K-means++ مقایسه می‌کنیم. نتایج مربوط به مقایسه با انتخاب مراکز تصادفی در جدول ۳ و نتایج مقایسه با روش K-means++ در جدول ۴ آمده است. با توجه به این‌که نتایج هر اجرای این دو روش با هم متفاوت است هر روش را ۱۰۰ مرتبه اجرا کردیم و میانگین، انحراف معیار و مقدار بیشینه معیار F را گزارش کردیم. با فرض

نرمال بودن توزیع می‌توان توزیع تجمعی نتایج حاصل از این دو روش را تخمین زد و احتمال آن‌که نتایج حاصل از روش پیشنهادی ما از نتایج دو روش دیگر بهتر باشد را محاسبه کرد که این احتمال را نیز به عنوان یکی از معیارهای عملکرد روش پیشنهادی گزارش کردیم که در ردیف پنجم جدول نشان داده شده است.

جدول ۳- مقایسه نتایج به دست آمده با K-means با مراکز تصادفی

مجموعه دادگان	20 newsgroups	TDT2	Reuters 21578
میانگین معیار F	0.624	0.726	0.381
انحراف معیار	0.022	0.010	0.010
بیشینه معیار F	0.652	0.751	0.405
معیار F روش پیشنهادی	0.635	0.814	0.499
$p(x \leq X)$	69.09%	100%	100%

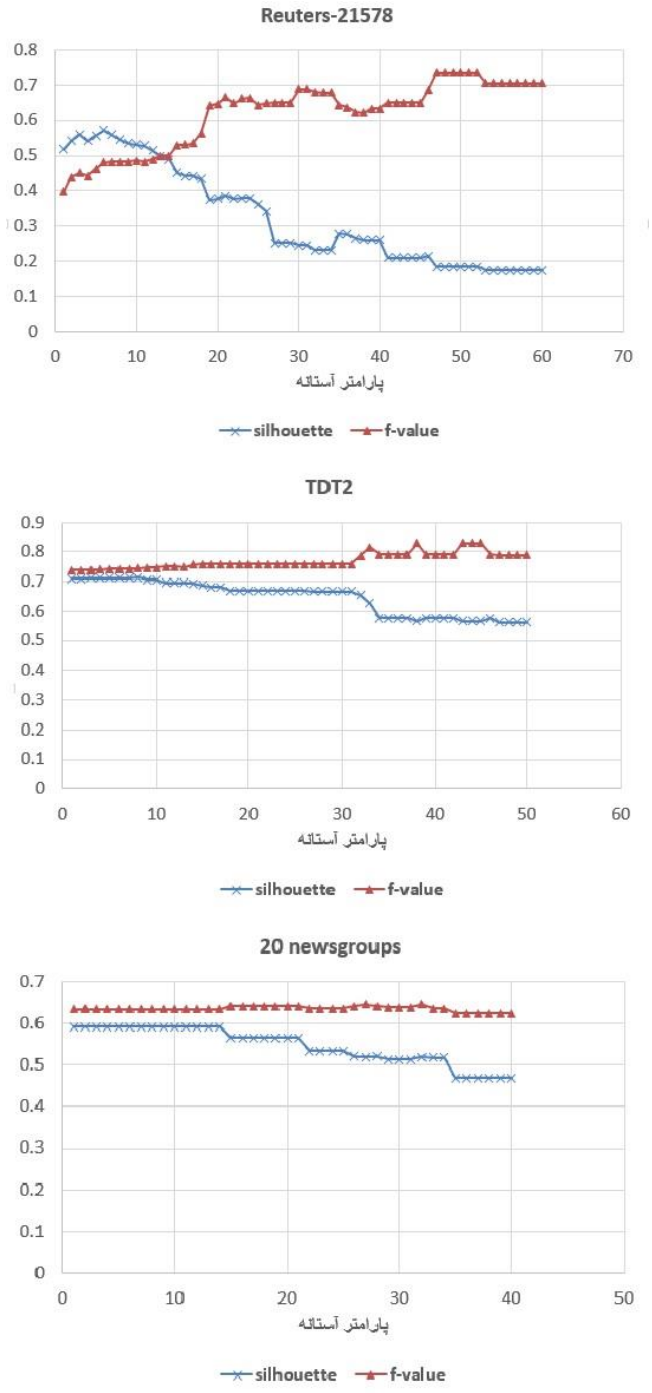
جدول ۴- مقایسه نتایج به دست آمده با K-means++

مجموعه دادگان	20 newsgroups	TDT2	Reuters 21578
میانگین معیار F	0.624	0.726	0.382
انحراف معیار	0.020	0.009	0.009
بیشینه معیار F	0.650	0.749	0.405
معیار F روش پیشنهادی	0.635	0.814	0.499
$p(x \leq X)$	71.82%	100%	100%

در سطر آخر جدول‌های ۳ و ۴ متغیر تصادفی x نماینده‌ی معیار F حاصل از اجرای دو روش مراکز تصادفی و K-means++ است و X بیان کننده‌ی نتایج روش پیشنهادی است. بنابراین $p(x \leq X)$ نشان دهنده‌ی احتمال بهتر بودن روش پیشنهادی نسبت به دو روش مذکور است. همچنین یکی از اطلاعات مهمی که از این جداول قابل برداشت است اعداد مربوط به انحراف معیار نتایج حاصل از انتخاب مراکز مختلف در الگوریتم K-means است که این اعداد نشان می‌دهند که روش خوشه‌بندی K-means مقدار زیادی به انتخاب مراکز اولیه حساس است.

۴-۴- تحلیل

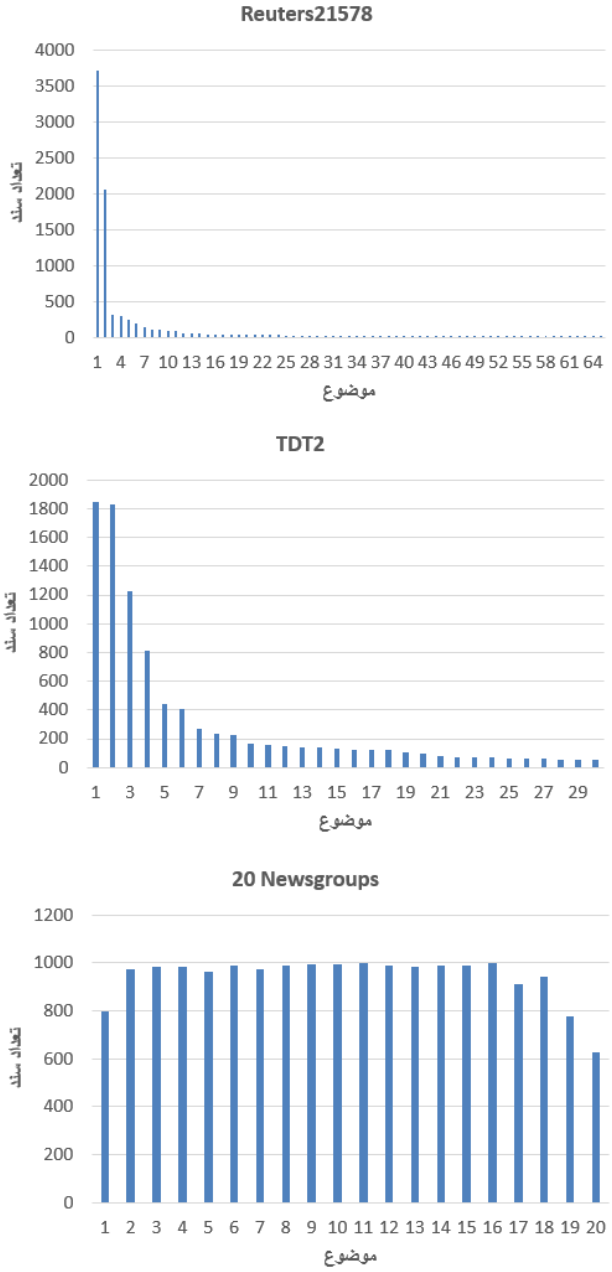
در این بخش از مقاله به بررسی و تحلیل نتایج به‌دست‌آمده می‌پردازیم. ابتدا با توجه به این‌که برای تعیین مقدار پارامتر آستانه در روش ارائه شده از یک معیار بی ناظر با عنوان Silhouette استفاده شده است در این قسمت ارتباط این معیار را با مقدار معیار F را بررسی می‌کنیم. در نمودارهای شکل ۳ این ارتباط بررسی شده است، در هر نمودار مقدار معیار F و Silhouette با افزایش میزان پارامتر آستانه برای هر یک از مجموعه‌های دادگان آورده شده است. همان‌طور که مشخص است تغییرات این دو معیار با یکدیگر مرتبط بوده و این شکل نشان می‌دهد که استفاده از روش گفته شده در بخش ۳-۳ منطقی بوده و نقطه‌ای که نمودار Silhouette بیشترین شکستگی را دارد محل مناسبی برای انتخاب این پارامتر است. تنها این ارتباط در مجموعه دادگان 20 newsgroups با کمی مشکل روبرو است که دلیل این مشکل و کاهش کیفیت نتایج در این مجموعه دادگان را در ادامه بررسی می‌کنیم.



شکل ۳- روند تغییرات Silhouette و معیار F

به منظور تحلیل نتایج به دست آمده در مجموعه‌های دادگان مختلف، به بررسی ویژگی‌های هریک از مجموعه دادگان پرداختیم. در شکل ۴ هر نمودار نشان دهنده‌ی توزیع اسناد در موضوعات مختلف در داده‌های برچسب خورده است. همان‌طور که ملاحظه می‌کنید توزیع داده‌ها در مجموعه‌های دادگان Reuters21578 و TDT2 به صورت توزیع توانی است به شکلی هر چه جلو می‌رویم تعداد اسناد هر دسته به شدت کاهش یافته و اکثر اسناد در چند دسته‌ی موضوعی اول قرار دارند. این در حالی است که در مجموعه دادگان 20 newsgroups این توزیع تقریباً یکنواخت بوده و تعداد اسناد موجود در دسته‌ها به هم نزدیک است. با توجه به این نکته که عمل خوشه‌بندی در مجموعه‌های دادگانی که توزیع توانی دارند بسیار مشکل‌تر از مجموعه دادگان با توزیع یکنواخت است [۳۵]، روش پیشنهادی ما توانسته است بهبود زیادی را در

دو مجموعه دادگان Reuters21578 و TDT2 که دارای توزیع توانی هستند داشته باشد و با توجه به آسان‌تر بودن عمل خوشه‌بندی در مجموعه دادگان 20 newsgroups که توزیع یکنواختی دارد روش پیشنهادی ما نتوانسته است بهبود چشم‌گیری نسبت به رقیب‌ها داشته باشد، در نتیجه روش ما را می‌توان مناسب برای مجموعه دادگان با توزیع توانی دانست.



شکل ۴- توزیع داده‌ها در موضوعات هر یک از مجموعه دادگان

همان‌طور که در بخش ۳-۲ گفته شد افزایش مقدار پارامتر آستانه باعث کاهش تعداد خوشه‌های نهایی می‌شود. از این رو تعداد خوشه‌های نهایی حاصل از روش پیشنهادی پس از تعیین مقدار مناسب برای پارامتر آستانه از تعداد کل موضوعات کمتر است. برای تکمیل نتایج و مقایسه کامل‌تر، تعداد خوشه‌هایی که از روش پیشنهادی ما به دست آمده است را به عنوان تعداد خوشه به الگوریتم LDA داده و آن را اجرا می‌کنیم. نتایج این مقایسه در جدول شماره ۵ مشاهده می‌شوند. همان‌طور که مشخص است در هر سه مجموعه دادگان میزان معیار F در روش پیشنهادی ما بیشتر از روش LDA با تعداد خوشه‌های برابر است. این امر نشان می‌دهد نتایج بهتر روش پیشنهادی ما نسبت به روش LDA تنها مربوط به کاهش تعداد خوشه‌ها نیست و کارایی روش ما حتی با تعداد خوشه‌های برابر نیز از روش LDA بهتر است.

جدول ۵- مقایسه نتایج با روش LDA با تعداد خوشه برابر

مجموعه دادگان	20 newsgroups	TDT2	Reuters 21578
تعداد خوشه‌ها	17	19	19
معیار F روش LDA	0.580	0.765	0.580
معیار F روش پیشنهادی	0.635	0.814	0.647

۵- نتیجه‌گیری و جمع‌بندی

در این مقاله روش جدیدی برای مسئله تشخیص موضوع ارائه کردیم که این روش با بهره‌گیری از توزیع موضوع حاصل از LDA اسناد و ارائه روش جدیدی برای انتخاب مراکز اولیه برای الگوریتم K-means عمل کرده و باعث بهبود کیفیت نتایج می‌شود. در روش ارائه شده ابتدا توزیع موضوع حاصل از LDA را محاسبه می‌کنیم و از آن برای محاسبه فاصله در روش خوشه‌بندی K-means و همچنین در الگوریتم ارائه شده برای پیدا کردن مراکز اولیه استفاده می‌کنیم. سپس با داشتن توزیع موضوع حاصل از LDA اسناد روشی را ارائه کردیم که هدف آن پیدا کردن مراکز اولیه مناسب برای الگوریتم K-means است. همچنین با توجه به این نکته که در الگوریتم ارائه شده برای پیدا کردن مراکز اولیه مناسب، پارامتر آستانه‌ای وجود دارد که باید مقدار مناسبی را برای آن پیدا کنیم، در این مقاله روشی بدون در نظر گرفتن داده‌های برچسب خورده برای تعیین کردن این مقدار ارائه شده است و روش ارائه شده برای مسئله تشخیص موضوع در مجموع بی ناظر است.

پس از ارائه روش پیشنهادی به انجام آزمایش‌ها بر روی سه مجموعه دادگان استاندارد در این مسئله پرداختیم و نتایج را با روش LDA برای تشخیص موضوع و دو روش پیدا کردن مراکز اولیه برای الگوریتم K-means یعنی انتخاب تصادفی مراکز و روش K-means++ مقایسه کردیم. نتایج نشان دهنده‌ی افزایش ۳۹ و ۹ درصدی در معیار F در زمان استفاده از روش پیشنهادی نسبت به روش LDA به ترتیب در مجموعه دادگان‌های Reuters21578 و TDT2 است و در مجموعه دادگان 20 newsgroups شاهد بهبود چشم‌گیری نبودیم. همچنین روش پیشنهادی در مقایسه با دو روش تعیین مراکز اولیه تصادفی و K-means در دو مجموعه دادگان Reuters21578 و TDT2 همیشه عملکرد بهتری داشته و از مقدار بیشینه آزمایش‌های این روش به‌اندازه‌ی قابل‌توجهی بیشتر است، همچنین در مجموعه دادگان 20 newsgroups روش پیشنهادی با احتمال نزدیک به ۷۰ درصد عملکرد بهتر نسبت به دو روش دیگر دارد. نتایج به دست آمده از آزمایش‌های ما نشان می‌دهد که الگوریتم K-means به انتخاب مراکز اولیه مناسب بسیار حساس بوده روش ما می‌تواند نقش مهمی در کیفیت آن داشته باشد، همچنین در این مقاله با توجه به تفاوت در نتایج در مجموعه دادگان مختلف به بررسی و تحلیل این تفاوت‌ها پرداختیم. در کارهای آینده نیز تمرکز ما بر روی بهبود کیفیت با استفاده از روش‌های خوشه‌بندی دیگر و نیز فراتر رفتن از مسئله تشخیص موضوع و بررسی و گسترش روش ارائه شده در مسائل دیگر مرتبط با خوشه‌بندی خواهد بود.

مراجع

[1] J. Allan, "Introduction to topic detection and tracking," in Topic detection and tracking, Springer, Boston, MA, 2002, pp. 1-16.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[3] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Systems with Applications, vol. 40, no. 1, pp. 200-210, 2013.

[4] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," In Proceedings of the 21th annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 28-36.

[5] J. Weng, and B.-S. Lee, "Event Detection in Twitter.," ICWSM, vol. 11, pp. 401-408, 2011.

[6] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia, "Event detection and popularity prediction in microblogging," Neurocomputing, vol. 149, pp. 1469-1480, 2015.

[7] W. Zhang, T. Chen, G. Li, J. Pang, Q. Huang, and W. Gao, "Fusing cross-media for topic detection by dense keyword groups," Neurocomputing, vol. 169, pp. 169-179, 2015.

[8] D. Spina, J. Gonzalo, and E. Amigó, "Learning similarity functions for topic detection in online reputation monitoring," In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 527-536.

[9] X. Wang, and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 424-433.

[10] L. Hong, and B. D. Davison, "Empirical study of topic modeling in twitter," In Proceedings of the first workshop on social media analytics, 2010, pp. 80-88.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," In Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004, pp. 487-494.

[12] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing Microblogs with Topic Models," In Proceedings of Fourth International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[13] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, 2009, pp. 248-256.

[14] J.F. Yeh, Y.S. Tan, and C.H. Lee, "Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation," Neurocomputing, vol. 216, pp. 310-318, 2016.

[15] W. Sriurai, "Improving text categorization by using a topic model," Advanced Computing, vol. 2, no. 6, p. 21, 2011.

[16] J. MacQueen, and others, "Some methods for classification and analysis of multivariate observations," In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, vol. 1, pp. 281-297.

[17] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," Theoretical Computer Science, vol. 38, pp. 293-306, 1985.

[18] D. Arthur, and S. Vassilvitskii, "k-means++: The advantages of careful seeding," In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp. 1027-1035.

[19] T. Onoda, M. Sakai, and S. Yamada, "Careful seeding method based on independent components analysis for k-means clustering," Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 1, pp. 51-59, 2012.

[20] J. A. Hartigan, and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100-108, 1979.

[21] G. N. Lance, and W. T. Williams, "A general theory of classificatory sorting strategies II. Clustering systems," The computer journal, vol. 10, no. 3, pp. 271-277, 1967.

[22] C. C. Aggarwal, and C. K. Reddy, "Data clustering: algorithms and applications," CRC press, 2013.

[23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics, vol. 20, pp. 53-65, 1987.

[24] D. Cai, X. Wang, and X. He, "Probabilistic dyadic data analysis with local and global consistency," In Proceedings of the 26th annual international conference on machine learning, 2009, pp. 105-112.

[25] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," In Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 911-920.

[26] D. Cai, X. He, W. V. Zhang, and J. Han, "Regularized locality preserving indexing via spectral regression," In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007, pp. 741-750.

[27] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 12, pp. 1624-1637, 2005.

[28] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, "The TDT-2 text and speech corpus," In Proceedings of the DARPA Broadcast News workshop, 1999, pp. 57-60.

[29] W. Li, J. Joo, H. Qi, and S.-C. Zhu, "Joint Image-Text News Topic Detection and Tracking with And-Or Graph Representation," arXiv preprint arXiv:1512.04701, 2015.

[30] D. D. Lewis, "Reuters-21578 text categorization test collection, distribution 1.0," http://www. research. att. com/~ lewis/reuters21578. html, 1997.

[31] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," DTIC Document, 1996.

[32] C. Wartena, and R. Brussee, "Topic detection by clustering keywords," In Proceedings of the 19th International Workshop on Database and Expert Systems Applications, 2008, pp. 54-58.

[33] X. H. Phan, L. M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," In Proceedings of the 17th international conference on World Wide Web, 2008, pp. 91-100.

[34] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent dirichlet allocation in web spam filtering," In Proceedings of the 4th international workshop on Adversarial information retrieval on the web, 2008, pp. 29-32.

[35] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," Computer, vol. 32, no. 8, pp. 68-75, 1999.

سپهر آروین در سال ۱۳۹۶ در مقطع کارشناسی‌ارشد رشته مهندسی فناوری اطلاعات (IT) از دانشکده‌ی مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی دانشگاه تهران فارغ‌التحصیل شد. زمینه‌های پژوهشی مورد علاقه او پردازش هوشمند متن و زبان طبیعی، داده‌کاوی، بازیابی اطلاعات، شبکه‌های اجتماعی و متن کاوی می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

s.arvin@ut.ac.ir

علی ورداسبی در سال ۱۳۹۷ با درجه‌ی دکتری مهندسی نرم‌افزار از دانشگاه تهران فارغ‌التحصیل شد. زمینه‌های پژوهشی او پردازش زبان طبیعی و شبکه‌های اجتماعی هستند. به‌طور خاص، او در موضوع‌هایی شامل خلاصه‌سازی متن، تشخیص موضوع، بیشینه‌سازی تأثیر و تحلیل معنایی مقالاتی در همایش‌ها و مجلات داخلی و خارجی به چاپ رسانده است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

a.vardasbi@ut.ac.ir

هشام فیلی تحصیلات خود را در مقاطع کارشناسی و ارشد در گرایش نرم‌افزار دانشگاه صنعتی شریف و سپس مقطع دکتری را در گرایش هوش مصنوعی از همان دانشگاه به اتمام رساند. و از سال ۱۳۸۷ به عنوان عضو هیات علمی دانشکده‌ی مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی دانشگاه تهران مشغول فعالیت است. زمینه‌های علاقه ایشان پردازش هوشمند متن و زبان طبیعی، شبکه‌های اجتماعی و متن کاوی می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

hfaili@ut.ac.ir

آزاده شاکری تحصیلات خود را در مقاطع کارشناسی و ارشد در رشته مهندسی کامپیوتر دانشگاه صنعتی شریف و سپس مقطع دکتری را در رشته علوم کامپیوتر از دانشگاه ایلینویز در اربانا- شمپین به اتمام رساند. و هم اکنون به عنوان عضو هیات علمی دانشکده‌ی مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی دانشگاه تهران مشغول فعالیت است. زمینه‌های علاقه ایشان بازیابی اطلاعات، داده کاوی، پردازش هوشمند متن و زبان طبیعی، شبکه‌های اجتماعی و متن کاوی می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shakery@ut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۰۵/۲۲

تاریخ اصلاح: ۱۳۹۶/۰۶/۱۷

تاریخ قبول شدن: ۱۳۹۶/۰۶/۲۷

نویسنده مرتبط: سپهر آروین، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران.

---

[1]Topic Detection and Tracking
[2]Topic Detection
[3]Hard Clustering
[4]Latent Dirichlet Allocation
[5]Variational Methods
[6]Author-Topic
[7]Hypernym
[8]Independent Component