یک روش وزن‌دهی مبتنی بر موقعیت واژه جهت مشابهت سنجی اسناد

مریم اسدی لنگرودی[۱]، سید ابوالقاسم میرروشندل [۲]

[۱] دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات ، گروه کامپیوتر، پردیس دانشگاه گیلان ، رشت
maryam.as16@yahoo.com

[۲] استادیار دانشگاه، گروه کامپیوتر ، دانشگاه گیلان ، رشت
mirroshandel@gmail.com

چکیده

اندازه‌گیری میزان شباهت اسناد موجود در وب، از آن جهت دارای اهمیت است که دربسیاری از زمینه‌ها، مانند بازیابی‌اطلاعات، دسته‌بندی متون، خوشه‌بندی اسناد، سیستم‌های تشخیص تقلب و سرقت ادبی، خلاصه‌سازی متون، و سایر حوزه‌ها، نقش مهم و اساسی ایفا می‌کند. میزان درستی این تشخیص، می‌تواند کارایی، دقت و صحت هر یک از فرایندهای مذکور را بالا برد. در تمام روش‌های مشابهت‌سنجی اسناد، اصول کار، تشخیص شباهت سندها بر مبنای شناخت دقیق ویژگی‌های مشترک آنها است. از این رو بازنمود سند بر مبنای ویژگی های بارز آن بسیار با اهمیت است. روش‌های مهم مشابهت‌سنجی اسناد، شامل مدل‌های لغوی و معنایی مبتنی بر محتوا و مدل‌های مبتنی بر ساختار صفحه است. در روش‌های لغوی ،ویژگی اصلی یک سند، وزن واژه‌های آن است. بنابراین وزن‌دهی صحیح واژه، می‌تواند یک گام موثر درنمایش دقیق ویژگی‌های بارز اسناد باشد.

در این مقاله، هدف اصلی، ارایه روشی بهبودیافته در وزن‌دهی واژه، برای تعیین میزان شباهت لغوی اسناد متنی است. روش وزن‌دهی پیشنهادی بر مبنای طرح TD - IDF و با تاکید بر اهمیت بیشتر واژه‌های خطوط ابتدایی اسناد، توانسته است، دقت و فراخوانی را در دسته‌بندی و میزان صحت را در خوشه‌بندی اسناد مجموعه TDT5 افزایش دهد.

کلمات کلیدی

شباهت اسناد، معیار شباهت، وزن‌دهی واژه، خوشه‌بندی اسناد، دسته‌بندی اسناد، بازیابی‌اطلاعات، شباهت‌سنجی لغوی، TD - IDF

۱- مقدمه

در بیشتر رهیافت‌های بازیابی‌اطلاعات توسط موتورهای‌جستجو، مبنای جستجو، تعیین میزان شباهت اسناد با پرس‌وجو و سپس رتبه‌بندی اسناد براین اساس است[۳-۱]. در حوزه داده‌کاوی و دسته‌بندی و خوشه‌بندی اسناد نیز مبنای کار، روش‌هایی است که میزان شباهت سندها را با دقت بالا بررسی می‌کند[۶-۴]. مشابهت‌سنجی اسناد، درحوزه‌های دیگر نیز چون خلاصه‌سازی متن، سیستم‌های پرسش‌وپاسخ، ترجمه ماشین، تشخیص تقلب و سرقت ادبی، نقش مهم ایفا می‌کند [۷].

شیوه‌های شباهت سنجی اسناد می تواند مبتنی برمحتوا و مبتنی برساختار صفحه باشد [۸,۹] . در شیوه‌های مبتنی بر محتوا سه دسته‌بندی مهم شباهت‌سنجی لغوی، معنایی و ترکیبی وجود دارد. روش‌های لغوی به بررسی میزان هم‌پوشانی اسناد و تعداد واژه‌های مشترک میان آنها می‌پردازند. روش‌های معنایی اسناد را از نظر مفهوم و معنی واژه‌ها بررسی می‌کند شیوه‌های ساختاری نیز به بررسی مشابهت ساختاری صفحات، مانند ارجاعات و لینک‌های ورودی و خروجی یکسان در اسناد می‌پردازند. در تمام این روش‌ها، پس از پیش‌پردازش‌های لازم، استخراج ویژگی‌های اسناد صورت می‌گیردو الگوریتم‌های سنجش

- مبنای معیارهای تعیین شباهت، میزان مشابهت ویژگی‌های اسناد را اندازه‌گیری می‌کنند [۱۲-۱۰].
- در تمام روش‌های مذکور، تشخیص دقیق ویژگی‌های با اهمیت اسناد، مدل نمایش سند، الگوریتم‌های مشابهت‌سنجی و معیارهای تعریف شباهت حائز اهمیت است و بهبود در هر یک ، می تواند منجر به تشخیص دقیق تر شباهت اسناد شود.
- در مدل شباهت سنجی لغوی اسناد به صورت برداری از وزن واژه‌های سند به عنوان ویژگی‌های اصلی نمایش داده می شوند. معروف ترین طرح وزن دهی در این روش، TF-IDF است. این طرح بر مبنای تعداد تکرار یک واژه در یک سند وکل اسناد مجموعه متون به هر واژه سند یک درجه اهمیت تخصیص می‌دهد. و جایگاه واژه در سند لحاظ نمی‌شود. این طرح دارای معایبی است. نخست این که یک طرح موردی است و زیر بنای مفاهیم ریاضی ندارد و دوم این که ابعاد وسیع و اندازه بزرگ مجموعه واژگان در یک مجموعه اسناد، حجم محاسبات را بسیار زیاد می‌نماید. این طرح، جایگاه واژه را در سند لحاظ نمی کند و تمام واژه‌ها در سند از فرمول وزن‌دهی یکسان استفاده می‌کند. همچنین تنها نمایان‌گر اهمیت لغوی یک واژه است و معانی واژه‌ها را در نظر نمی‌گیرد.همچنین دارای معایبی چون عدم توجه به مدل توزیع واژه در سند ، مدل توزیع بین کلاسی، اهمیت طول و محل قرارگیری واژه وسایر موردها است . مطالعات زیادی در این زمینه انجام گرفته است که به طور موردی تعدادی از آنها به شرح زیربیان شده است.

- در [۱۳]، بااستفاده از طرح TF-IDF و الگوریتم KNN ، تعداد ۷۵۰۰مقاله خبری کشور اندونزی با ۱۵ دسته موضوعی از پیش‌تعریف‌شده، دسته‌بندی شدند و صحت این دسته‌بندی حدود ۹۸٪ بیان شده است.در [۱۴] طرح وزن‌دهی TF-IDF ,LSI, MULI-WORD بر روی اسناد به زبان های چینی و انگلیسی مورد مطلعه قرار گرفت که در دسته‌بندی اسناد انگلیسی، این طرح در مرتبه دوم صحت دسته‌بندی قرار گرفت. در [۱۵] لن و همکاران، عملکرد ۹ طرح سنتی از جمله Binary، TF، ITF، TF - IDF، TF - IDF - LOG TF - IDF، TF IDF Prob، و TF - Chi را در مجموعه داده‌های های بزرگ با الگوریتم دسته بندی SVM مقایسه کردند. آن‌ها نتیجه گیری نمودند که TF - Chi و TF طرح‌هایی با بهترین عملکرد است. همچنین یک طرح خاص وزن دهی به نام TF - RFرا پیشنهاد دادند که از فراوانی مربوط یک واژه استفاده می کند و ادعا کردند که طرح جدید آن‌ها، دارای عملکرد بهتری نسبت به طرح‌های دیگر است . در[۱۶]، مشکل بارز برای طرحTF-IDF عدم توجه به روابط میان واژه‌های هم‌معنی و عدم شناخت روابط کلمات ارجاعی به ضمیر، مطرح شد و به‌این منظور از ترکیب پیکره‌متون و روش‌های معنایی با طرح فوق بهره گرفته شد. در[۱۷]، یک طرح بهبودیافته TF-IDF با تمرکز بر اهمیت مکانی واژه و همچنین طول یک واژه بیان‌شده‌است. این طرح دارای دقت و فراخوانی بالاتری نسبت به طرح اولیه است. در[۱۸]بیان شده است که طرح TF-IDF، در دسته‌بندی اسناد، توزیع بین‌کلاس‌ها و داخل‌کلاسی واژه یا عبارت t_i را در نظر نمی‌گیرد و مشکل اصلی را در قسمت IDFطرح بیان کرده است. سپس یک الگوریتم تکرارشونده جهت بهبود طرح معرفی شد. در [۱۹]یک طرح بهبودیافته بر مبنای اصلاح عامل TF، ارایه شد. طرح پیشنهادی، اهمیت مدل توزیع یک واژه را در یک سند بررسی‌نمود و اهمیت واژگان را در توزیع یکنواخت و غیر متراکم در سند بیان کرد.
- وزن‌دهی دقیق واژه‌ها می‌تواند ویژگی های مهم اسناد را نمایش دهد. در این مقاله ابتدا مفاهیم اولیه روش‌های شباهت‌سنجی لغوی مبتنی بر محتوا و روش های وزن دهی ، بیان می‌شود و سپس، یک روش جدید که با تغییر مدل وزن‌دهی واژه بر مبنای اهمیت بیشتر واژه‌های خطوط ابتدایی سند توانسته است، بهبود قابل توجهی در کارایی، دقت و فراخوانی دسته‌بندی و صحت خوشه‌بندی اسناد ایجاد نماید، پیشنهاد می‌گردد و حاصل اجرای این روش بر روی اسناد مجموعه متونTDT5 مورد مطالعه و ارزیابی قرار می‌گیرد.

۲- نمایش سند در مشابهت سنجی لغوی

مشابهت سنجی لغوی، اسناد را بر اساس میزان هم‌پوشانی آن‌ها بررسی می‌کند. در این روش، اندازه‌گیری میزان شباهت، بر اساس تعداد واژه‌های مشترک موجود در اسناد است و مفهوم و معنی واژه در زبان طبیعی حائز اهمیت نیست. ویژگی بارز سند، واژه های با وزن بالای آن است.

در این روش، برای نمایش سند از مدل فضای برداری سالتون استفاده می‌شود. هر سند به صورت کیسه‌ای از کلمات دریک فضای برداری، در نظر گرفته می‌شود و توالی و جایگاه واژه در سند، لحاظ نمی‌گردد. ابتدا پیش‌پردازش و عمل قطعه بندی اسناد صورت می‌گیرد و واژه‌های سند یا همان توکن‌ها استخراج می‌شوند. یک توکن می‌تواند یک واژه یا یک عبارت یا یک ngram از واژه‌های متوالی باشد و لزوما یک کلمه درفرهنگ لغات یک زبان طبیعی نیست. در این مقاله یک واژه به عنوان یک توکن در نظر گرفته شده است. وزن هر واژه، براساس مدل‌های وزن‌دهی محاسبه می‌گردد و یک سند به صورت برداری از وزن واژه‌های آن دریک فضای برداری به ابعاد تعداد واژه‌های مجموعه نمایش داده می‌شود[۲۰]. در مجموعه‌ی اسناد D، $V=\{t1,t2,\dots tm\}$ مجموعه‌ای از واژه‌های متمایز است، که در آن، ti یک واژه است. مجموعه V معمولا به عنوان مجموعه واژگان نامیده می‌شود، و m اندازه یا تعداد واژه‌های آن است.

وزن هر واژه ti در سند $dj \in D$، برابر با wij است که اهمیت آن واژه را در آن سند، بیان می‌کند. برای واژه‌ای که در سند dj نمایان نمی‌شود، wij=0 خواهد بود. بنابراین هر سند ،dj با یک بردار واژه به صورت فرمول(۱) نمایش داده می‌شود:

$$d_j=\{w_{1j},w_{2j},\dots,w_{mj}\} \quad (۱)$$

که درآن d_j یک بردار سند در فضای m بعدی و wij وزن واژه است.

۲-۱- پیش پردازش اسناد برای نمایش

اسناد در وب، به صورت صفحات حاوی کدها و تگ‌های HTML است. برای تبدیل اسناد متنی به صورت برداری از واژه‌ها، ابتدا اعمال پیش‌پردازش روی آن‌ها صورت می‌پذیرد. ابتدا محتوای تگ‌ها از آن‌ها استخراج می‌شود و بدنه اصلی متن سند حاصل می‌گردد سپس تگ‌ها حذف می‌شوند. در بسیاری

از الگوریتم‌ها، به جهت اهمیت بعضی از تگ ها مانند تگ عنوان و یا سایر تگ‌های مورد اهمیت در مدل‌های شباهت ساختاری، شاخص‌های ویژه‌ای برای آن‌ها در نظرگرفته می‌شود. در مرحله بعد، عمل قطعه‌بندی یا توکن‌بندی سند صورت می‌پذیرد. و واژه‌های سند استخراج می‌گردد. با توجه به ابعاد بزرگ داده‌ها، از پردازش‌هایی جهت کاهش ابعاد واژه‌ها استفاده می‌شود، مانند حذف علایم و اعداد. یکی از پردازش‌های مهم در این زمینه، حذف کلمات توقف است. کلمات توقف در یک زبان طبیعی واژه‌هایی است پرتکرار که بار معنایی خاصی ندارد. به عنوان مثال، در زبان انگلیسی کلماتی چون"And" و"The" کلمات توقف به‌شمار می‌آیند. سپس هر توکن، ریشه‌یابی‌می‌شود و ریشه یا مصدر اصلی واژه، جایگزین آن می‌گردد. بسیاری از مدل‌ها، از روش‌های دیگرین نیز برای کاهش ابعاد بهره می‌جویند.

پس از پیش پردازش اسناد، عملیات‌وزن‌دهی به توکن‌ها بر اساس انواع طرح‌های وزن‌دهی،انجام می‌شود و بردار سند واژه ایجاد می‌گردد.

۲-۲- روش‌های وزن‌دهی واژه

وزن‌دهی واژه، یکی از مهم‌ترین مراحل در نمایش یک سند بر اساس اهمیت ویژگی‌های آن است. این امر می‌تواند ویژگی‌های شاخص یک سند را نمایان کند. عمل شباهت‌سنجی اسناد بر اساس تطبیق وزن واژه‌ها به عنوان ویژگی‌ها، انجام می گیرد. در روش‌های مختلف شباهت‌سنجی اسناد از طرح های گوناگون وزن‌دهی استفاده شده است.

یک طرح مهم وزن‌دهی در مدل نمایش برداری اسناد، طرح مشهور TF – IDF است. این طرح،یک عدد آماری ارایه می‌دهد که نشان دهنده درجه اهمیت یک واژه در یک سند از مجموعه اسناد یا پیکره متنی است[۲۱,۲۲].

TF ، معیاری است که ارزش آن، تعداد تکرار یک واژه در یک سند واحد است. در این حالت، واژه‌ی پرتکرار در یک سند، جزء با اهمیت‌ترین واژه‌های آن سند، محسوب می‌شود. از سوی دیگر اگر آن واژه، درسایر اسناد مجموعه و یا پیکره‌متنی نیز به کار رفته باشد، یک واژه شاخص برای آن سند محسوب نمی‌گردد. معیار معکوس تکرار سند یا IDF می‌تواند این حالت را متعادل نماید. این معیارعدد متعادل شده تعداد اسناد حاوی یک واژه را نسبت به کل اسناد مجموعه نشان می‌دهد.فرمول (۲) محاسبه این معیار را نشان می‌دهد[۲۳]:

$$IDF_i = LOG\left(\frac{N}{df_i}\right) \qquad (۲)$$

که درآن N ، تعداد کل اسناد مجموعه اسناد یا پیکره متنی و dfi تعداد سندهایی است که واژه ti، در آن‌ها وجود دارد.

در این طرح وزن واژه ti، در سند dj به صورت فرمول(۳)، خواهد بود که wij وزن واژه است:

$$W_{ij} = TF_{ij} \cdot IDF_i \qquad (۳)$$

در این طرح، واژهای دارای بیشترین اهمیت و ویژگی شاخص یک سند است که در آن سند بیشترین تکرار و در سایر اسناد مجموعه کمترین رخداد را داشته باشد.

در این مقاله با تغییر در مدل وزن‌دهی واژه‌ها، بر اساس این طرح ، روشی جدید با تاکید بر اهمیت واژه‌های خطوط اولیه اسناد مطرح شده است. مدل پیشنهادی، اهمیت موقعیت و جایگاه یک واژه را در یک سند در نظر گرفته و توانسته است مشابهت سنجی دقیق‌تری را حاصل کند.

۲-۳- معیار مشابهت سنجی

تعیین میزان شباهت یا فاصله بردارهای اسناد، نیازمند استفاده از معیارهایی برای اندازه‌گیری است. معیارهای زیادی برای این اندازه‌گیری وجود دارد بعضی ازاین معیارها، میزان شباهت و بعضی میزان فاصله را محاسبه می‌کنند که قابل تبدیل به یکدیگر است. از مهمترین آن‌ها در بررسی شباهت‌سنجی متون با در نظرگرفتن توکن‌های تک‌واژه‌ای، معیار شباهت کسینوسی، ضریب‌دایس، ضریب‌جاکارد، فاصله اقلیدسی، فاصله منهتن و معیار کالبک لیبلر است.

وقتی اسناد، به صورت برداری از ویژگی‌ها نمایش داده‌می‌شوند، مقدارکسینوس زاویه بین بردارها می‌تواند معیار اندازه‌گیری میزان شباهت آن‌ها باشد. این اندازه در بازه [۰,۱] است. در صورتی که دو بردار منطبق بر هم و یا کاملا شبیه به هم باشند، این عدد ۱ خواهد بود، شباهت کسینوسی از مشهورترین معیارهای تشخیص شباهت متون در حوزه‌های بازیابی‌اطلاعات و خوشه‌بندی اسناد است[۲۴]. فاصله اقلیدسی نیز یک معیار متریک مطرح و موفق تعیین میزان فاصله است که قابلیت تبدیل به معیار شباهت را دارد. این معیار پیش‌فرض، در الگوریتم‌های خوشه بندی K - Means استفاده می‌شود.

۳- مبنای روش پیشنهادی

در حوزه خلاصه سازی متن، امتیاز دهی به جملات، از اهمیت بالایی برخوردار است و یکی از مدل‌های مبنایی این حوزه به شمار می‌رود. در روش‌های خلاصه‌سازی متون، به دلیل ابعاد وسیع ویژگی‌های متن، مبنای کار ، استخراج ویژگی‌های مهم‌تر و بارز سند است. مدل‌های مبتنی بر موضوع ، مبتنی بر اهمیت جمله ومبتنی بر موقعیت واژه از جمله روش‌های با اهمیت در خلاصه‌سازی خودکار متون محسوب می‌شوند. در مدل‌های مبتنی بر موقعیت واژه، اهمیت واژه‌هایی که در موقعیت‌های مکانی خاص از متن قرار دارند، بالاتر است. به‌طور مثال درمقالات خبری و روزنامه‌های الکترونیکی، جمله اول متن، دارای رتبه بالاتری نسبت به سایر جملات است و یاجمله اول هر پاراگراف در مقالات خبری و جملات آخر در مقالات علمی از اهمیت بیشتری برخوردارند[۲۷-۲۵]. بر این مبنا، روش پیشنهادی ارائه می‌گردد.

٤- روش وزن‌دهی پیشنهادی TF – IDF_P

در روش پیشنهادی TF – IDF_P ، فرض بر این است که خطاهو جملات اولیه یک سند، می‌تواند حاوی چکیده مطالب کل سند و دارای بیشترین اهمیت محتوایی آن سند باشد. به طور مثال، در مقالات علمی خطوط اول، حاوی چکیده مقاله و در مقالات خبری، حاوی عناوین اصلی خبر است. از این‌رو، هنگام وزن‌دهی توکن‌ها بر اساس این طرح به توکن‌هایی که جزء ۲۵۰ کاراکتر نخست یک سند باشند و در خطهای اولیه سند قراربگیرند، ضریب ۵ برابر وزن حاصل از طرح اولیه TF - IDF، اختصاص می‌یابد و سایر واژه‌ها از ضریب ۱ این طرح برخوردار خواهندبود. فرمول (۴) الگوریتم اجرایی این روش را نشان می‌دهد:

IF POS(t_{ij}) <=250 THEN
$W_{ij} = 5 * TF_{ij}.IDF_i$ (۴)
Else
$W_{ij} = TF_{ij}. IDF_i$

که تابع POS موقعیت مکانی واژه t_{ij} را در سند، بیان می‌کند.

بر این اساس، دو سندی که در خطوط اولیه خود دارای واژه‌های مشترک باشند، امکان شباهت بیشتری با یکدیگر دارند. این مدل وزن‌دهی، توانسته است در بهبود عملکرد خوشه‌بندی و دسته‌بندی اسناد، تاثیرگذار باشد. به طور مثال اگر توکن "air"، درموقعیت محدوده ۲۵۰ کاراکتر اول یک سند، ۳ مرتبه تکرار شده باشد و در سایر مکان‌های سند، در مجموع ۶ مرتبه، رخ داده باشد، عدد تکرار این واژه در سند بر اساس روش پیشنهادی، عدد ۲۱ خواهد بود.در روش اولیه این عدد ۹ است.

٤-١- استخراج داده

در این روش، از اسناد حاوی متن به زبان انگلیسی مجموعه داده TDT5، استفاده شده‌است. لازم به‌ذکر است که فایل‌های این مجموعه، اسناد خبری است. از تعداد ۳,۳۲۳ فایل با ساختار HTML این مجموعه، تعداد۱,۷۶۳,۶۹۹، تگ <doc> به عنوان سند مجزا، استخراج شده است. این اسناد توسط عملیات پیش‌پردازش، به قالب ساختارمند تبدیل‌گردید .از این تعداد،۱,۲۵۸,۰۰۰ سند دارای تگ <keyword> که نمایان‌گر موضوع‌خبر است، انتخاب شد و سایر اسناد از مجموعه، خارج شدند. از تعداد مذکور سند،۵,۷۷۶ موضوع‌خگجبر منحصربفرد حاصل‌شد. به منظور کاهش‌ابعاد داده، تعداد ۱۰، keyword یا همان موضوع‌خبری پرتکرار انتخاب گردید و از هر موضوع خبری واحد به صورت تصادفی تعداد ۱۰۰۰ سند برگزیده شد. در نهایت، تعداد ۱۰,۰۰۰ سند، به عنوان مجموعه داده‌های این آزمایش در نظرگرفته شد.

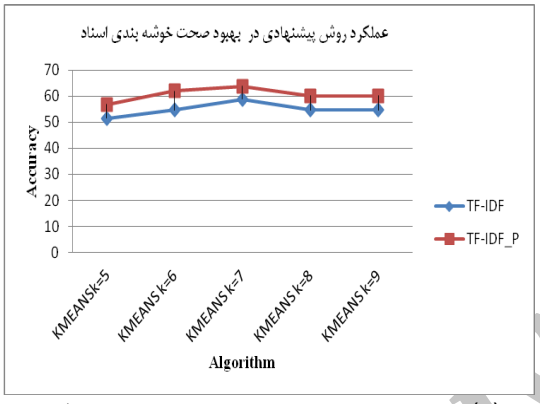
٤-٢- پیاده سازی و ارزیابی روش

پس‌از اجرای پیش پردازش اولیه و حذف کلمات توقف ، تعداد ۱,۶۱۰,۰۵۴ واژه از مجموعه اسناد به دست آمدلازم به ذکر است، ۶۶۷ کلمه توقف زبان انگلیسی موجود در سایت رایگان ranks مورد استفاده قرار گرفت.سپس با استفاده از الگوریتم ریشه یابی پورتز، ریشه هر واژه جایگزین آن گردید. در مرحله اصلی، وزن‌دهی به دو روش TF-IDF یعنی بدون در نظرگرفتن اهمیت محل قرارگیری واژه و به روش پیشنهادی TF-IDF_P، با در نظر گرفتن اهمیت ۵ برابری ۲۵۰ کاراکتر اول سند، وزن هر واژه محاسبه گردید. قبل از ساختن ماتریس سند واژه، به منظور کاهش و تعدیل زمان محاسباتی، بر اساس روش یادگیری، سه حد آستانه وزن، با مقادیر ۰.۰۶ و ۰.۰۷و۰.۰۸ برای انتخاب واژه‌ها در نظر گرفته‌شد که حد آستانه ۰.۰۷ قابل‌قبول‌ترین نتیجه را حاصل کرد.

دو ماتریس سند واژه از اسناد مجموعه و با وزن‌های بزرگتر یا مساوی کران۰.۰۷، برای واژه‌ها ایجاد گردید. یک ماتریس با اعمال روش وزن‌دهی TF-IDF و ماتریس دیگر با روش وزن‌دهی TF-IDF_P حاصل شد. TA، ماتریس سند واژه با وزن‌دهی عادی، باتعداد ۲,۸۰۰ سند و ۳۴ واژه و TW، ماتریس سند واژه با وزن‌دهی پیشنهادی و با تعداد ۳,۰۳۸ سند و ۶۵ واژه حاصل گردید.

خوشه‌بندی اسناد، یک روش متن کاوی بدون نظارت است که اسناد متنی را به‌طور موثر بر اسا میزان شباهت آن‌ها، سازماندهی می کند. روش‌های خوشه‌بندی مبتنی بر چهار مفهوم اصلی، مدل نمایش داده‌ها، معیار شباهت، روش خوشه‌بندی، و الگوریتم خوشه‌بندی است.از آنجا که الگوریم‌های خوشه‌بندی تفکیکی، مانند K - means نسبت به الگوریتم‌های خوشه‌بندی سلسله‌مراتبی، برای بررسی مجموعه داده‌های سندی بزرگ، مناسب‌تر است و نیاز به محاسبات کمتری نیازدارد[۱۹]، برای پیاده‌سازی از این الگوریتم استفاده شده است[۲۸].

در پیاده‌سازی روش پیشنهادی ، با استفاده از نرم‌افزار ،Weka از الگوریتم خوشه بندی تفکیکی K - means با معیار شباهت فاصله اقلیدسی در ۵ مرحله با در نظرگرفتن تعداد خوشه‌های ۵,۶,۷,۸و۹، استفاده شده است.،که نمودار مقایسه‌ای آن در(شکل ۱) قابل مشاهده است:



شکل(۱) : ارزیابی عملکرد TF-IDF_P در صحت خوشه‌بندی

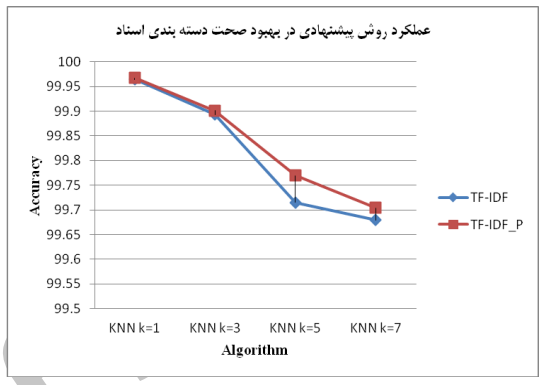
همان‌گونه که در (شکل ۱) مشخص است، روش وزن‌دهی پیشنهادی در تعداد خوشه‌های مختلف انتخابی دارای صحت بیشتر نسبت به روش وزن‌دهی بدون در نظرگرفتن تعداد کارکترهای اول بوده است و بهبود اجرای این روش، کاملاً در شکل نمایان است.

در مرحله بعد ،از ارزیابی دیگر به این منظور استفاده شد. در مجموعه اسناد، می‌توان براساس تگ‌های موردنظر یا با توجه به مشخصه‌های خاص سند، آن‌ها را در دسته‌هایی دارای برچسب مشخص، دسته‌بندی نمود. این مدل یک روش با نظارت در داده کاوی محسوب می‌شود. از مهم‌ترین الگوریتم‌های دسته‌بندی اسناد، روش K نزدیکترین همسایه است. این مدل بسیار به مدل‌های خوشه‌بندی تفکیکی، مشابه است و اشیاء را با توجه به مشابهت با اشیاء مجاور آن محاسبه می‌کند. این الگوریتم در دسته‌بندی متون، دقت و فراخوانی قابل قبولی را ارائه می‌دهد[۲۹]. روش های وزن‌دهی واژه تاثیر بسزایی در دسته بندی اسناد دارد.

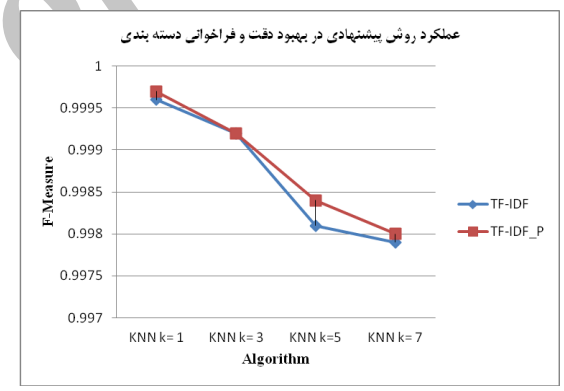
در این مرحله ، دو ماتریس طریق الگوریتم دسته‌بندی Kنزدیک‌ترین همسایه در چهار مدل تعداد همسایه‌های نزدیک ۵,۳,۱ و۷ و موضوع‌خبر به عنوان کلاس‌نهایی، در ۱۰ کلاس موضوعی، دسته‌بندی شدند. در هر مرحله دقت و فراخوانی دسته‌بندی و در نهایت میانگین معیار F – Measure و

Accuacy یا صحت، جهت ارزیابی عملکرد، محاسبه گردید که مطابق (شکل ۲) و (شکل۳)، بهبود حاصل از اجرای روش پیشنهادی مشهود است. سپس از طریق الگوریتم بیزین برای دسته‌بندی اسناد استفاده شد که در (شکل ۴) بهبود صحت مشخص شده است.

بهبود صحت دسته بندی بر حسب درصد، در تمام مراحل، به ویژه در شیوه اجرای الگوریتم مذکور با تعداد ۵ نزدیکترین همسایه از همه موارد بیشتر است. و این عدد از ۹۹.۷۱٪ به ۹۹.۷۷٪ افزایش یافته است. در تعداد همسایگان ۱و ۳ کمترین تغییر، حاصل‌شده‌است.

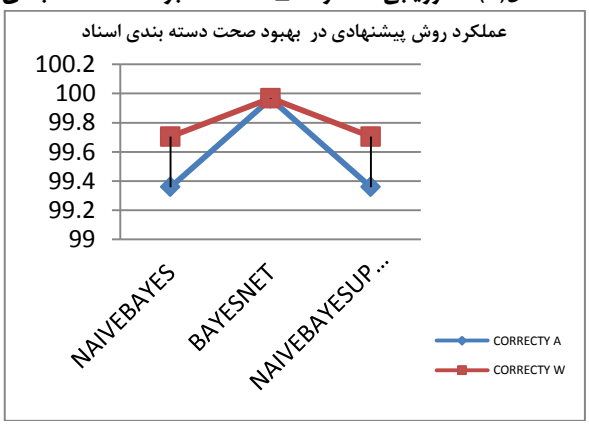


شکل(۲) : ارزیابی عملکرد TF-IDF_P در صحت دسته‌بندی



شکل(۳) : ارزیابی عملکرد TF-IDF_P در صحت دسته‌بندی

شکل(٤)- : ارزیابی عملکرد TF-IDF_P بر صحت دسته‌بندی



دقت و فراخوانی در دسته بندی با روش پیشنهادی بهبود یافته است و با تعداد۵ نزدیک‌ترین همسایه بهترین عملکرد را دارد. بالاتر است.

در اجرای الگوریتم K-means، میزان صحت در تعداد خوشه‌های ۸ و ۹، با هم برابر است. در مقایسه با روش اولیه، بهترین صحت خوشه‌بندی با تعداد ۶ خوشه، با بهبود حدود ۸٪ حاصل شده است.

۵- نتیجه

روش‌های وزن‌دهی واژه در محاسبه دقیق اهمیت واژه‌های اسناد به عنوان ویژگی‌های بارز برای نمایش سند،بسیار حائز اهمیت است. این امر تاثیر مستقیم بر عملکرد مشابهت سنجی اسناد و در نتیجه، دسته بندی و خوشه بندی متون دارد. در این مقاله، به بررسی طرح TF - IDF و عملکرد آن اشاره شد و بیان گردید که این طرح، اهمیت موقعیت مکانی یک واژه را در متن در نظر نمی‌گیرد، سپس طرح TF – IDF_P به عنوان یک روش بهبود یافته،پیشنهاد شد و عملکرد آن توانست بر صحت خوشه بندی و دقت و فراخوانی دسته بندی اسناد مجموعه متون TDT5 بهبودی قابل توجه حاصل کند.

ضمایم

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

TP تعداد موارد دسته‌بندی مثبت صحیح

TN تعداد موارد دسته‌بندی منفی صحیح

FP تعداد موارد دسته‌بندی مثبت نادرست

FN تعداد موارد دسته‌بندی منفی نادرست

$$F - \text{measure} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision} + \text{Recall}}$$

مراجع

[1] Singh, Jaskirat, and Mukesh Kumar. "A Meta Search Approach to Find Similarity between Web Pages Using Different Similarity Measures." Advances in Computing, Communication and Control. Springer Berlin Heidelberg, 150-160, 2011.

[2] Liu, Bing. "Information retrieval and Web search." Web Data Mining. Springer Berlin Heidelberg, 211-268, 2011.

[3] Roa-Valverde, Antonio J., and Miguel-Angel Sicilia. "A survey of approaches for ranking on the web of data." Information Retrieval 17, no. 4: 295-325, 2014.

[4] Huang, Anna. "Similarity measures for text document clustering."Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008.

[5] Li, Yanjun, Congnan Luo, and Soon M. Chung. "A parallel text document clustering algorithm based on neighbors." Cluster Computing 18.2: 933-948, 2015.

[6] Chim, Hung, and Xiaotie Deng. "Efficient phrase-based document similarity for clustering." Knowledge and Data Engineering, IEEE Transactions on 20.9: 1217-1229, 2008.

[7] Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." International Journal of Computer Applications 68.13: 13-18, 2013.
[8] Gipp, Bela. "Citation-based Document Similarity." Citation-based Plagiarism Detection. Springer Fachmedien Wiesbaden, 43-55, 2014.
[9] Liu, Hongyan, et al. "Measuring similarity based on link information: A comparative study." Knowledge and Data Engineering, IEEE Transactions on 25.12: 2823-2840, 2013.
[10] Zesch, Torsten, and Iryna Gurevych. "Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words." Natural Language Engineering 16.: 25-59, 2010.
[11] Chahal, Premjeet, Monika Singh, and Sudhakar Kumar. "Ranking of web documents using semantic similarity." Information Systems and Computer Networks (ISCON), 2013 International Conference on. IEEE, 2013.
[12] Machnik, Łukasz. "Documents Clustering Techniques." Annales UMCS Sectio AI Informatica 2.1: 401-411,2015.
[13] Hakim, Ari Aulia, et al. "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach." Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on. IEEE, 2014.
[14] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification." Expert Systems with Applications 38.3: 2758-2765, 2011.
[15] Lan, Man, et al. "A comprehensive comparative study on term weighting schemes for text categorization with support vector machines." Special interest tracks and posters of the 14th international conference on World Wide Web. ACM, 2005
[16] Huynh, Minh Chau, Pham Duy Thanh Le, and Trong Hai Duong,((2015), "Improved Vector Space Model TF/IDF Using Lexical Relations.", International Journal of Advanced Computer Research 5.21 334.P.
[17] Wang, Xingheng, et al.,,(2012), "Text clustering based on the improved TFIDF by the iterative algorithm.", Electrical & Electronics Engineering (EEESYM), 2012 IEEE Symposium on. IEEE..
[18] Wang, Xingheng, et al.,,(2012), "Text clustering based on the improved TFIDF by the iterative algorithm.", Electrical & Electronics Engineering (EEESYM), 2012 IEEE Symposium on. IEEE.
[19] Xia, Tian, et al,(2012), "An improved Global Weight Function of Terms based on Pearson's Chi-square statistics.", Information Science and Control Engineering 2012 (ICISCE 2012), IET International Conference on. IET.
[20] Handojo, Andreas, Adi Wibowo, and Yovita Ria. "Document Searching Engine Using Term Similarity Vector Space Model on English and Indonesian Document." Intelligence in the Era of Big Data. Springer Berlin Heidelberg, 165-173, 2015.
[21] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification." Expert Systems with Applications 38.3: 2758-2765, 2011.
[22] Hakim, Ari Aulia, et al. "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach." Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on. IEEE, 2014.
[23] Yoo, Jong-Yeol, and Dongmin Yang. "Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier." , 2015.
[24] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." Knowledge and Data Engineering, IEEE Transactions on 26.7: 1575-1590, 2014
[25] Krishna, RVV Murali, and Ch Satyananda Reddy. "Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis." Computational Intelligence in Data Mining—Volume 1. Springer India, 261-272, 2016.
[26] Kallimani, Jagadish S., K. G. Srinivasa, and B. Eswara Reddy. "Summarizing News Paper Articles: Experiments with Ontology-Based, Customized, Extractive Text Summary and Word Scoring." Cybernetics and Information Technologies 12.2: 34-50, 2012
[27] Osman, Ahmed Hamza, et al. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12.5: 1493-1502, 2012.
[28] Huang, Faliang, et al. "Clustering web documents using hierarchical representation with multi-granularity." World Wide Web 17.1: 105-126, 2014.
[29] Kumar, Jayant, Peng Ye, and David Doermann. "Learning document structure for retrieval and classification." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.