دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران
دانشگاه شهید بهشتی، دانشکده مهندسی برق و کامپیوتر ،تهران، ایران، ۱ تا ۳ اسفند ۱۳۸۵

TFCRF: روش جدید وزن دهی ویژگی مبتنی بر اطلاعات کلاس در حوزه طبقه بندی مستندات

مینا ملکی[۱]، احمد عبدالله زاده بارفروش[۲]

چکیده

وزن دهی ویژگی به عنوان یکی از تکنیک‌های پیش‌پردازش در حوزه طبقه بندی مستندات، نقش بسیار مهمی در دستیابی به شاخص بندی با کیفیت بالا و در نتیجه دستیابی به طبقه بندی کننده خوب مستندات ایفا می‌کند. در این مقاله یک روش جدید برای وزن‌دهی ویژگی به نام TFCRF خاص حوزه طبقه‌بندی مسستندات ارائه می شود که در آن برای وزن‌دهی ویژگی‌ها علاوه بر توجه به چگونگی توزیع آنها در مستندات مختلف و مستندات کل مجموعه به چگونگی توزیع آنها در طبقات مختلف نیز توجه شده است. نتایج شبیه سازی نشان دهنده بهبود قابل توجهی در کارایی الگوریتم طبقه بندی کننده SVM با بکارگیری روش وزن دهی ویژگی ارائه شده جدید TFCRF در مقایسه با سایر روش‌های متداول وزن دهی ویژگی پیاده‌سازی شده نظیر روش‌های مبتنی بر TF، روش‌های مبتنی بر IDF، روش‌های ترکیبی TFIDF و روش‌های خاص طبقه بندی بر روی مجموعه مستندات *inex* می‌باشد.

کلمات کلیدی

وزن دهی ویژگی، طبقه بندی مستندات، الگوریتم SVM، انتخاب ویژگی، بازیابی مستندات، متن کاوی.

TFCRF: A Novel Feature Weighting Method Based on Class Information in Text Categorization

Mina Maleki, Ahmad Abdollahzadeh Barforush

Abstract

Feature weighting which is one of the important preprocessing techniques in text categorization has a vital role in retrieving a high quality indexing and so in having a good text categorization system. In this paper, a new feature weighting method named TFCRF in text categorization domain is presented. The TFCRF considers both the distribution of the feature within different documents and its distribution within different categories for weighting a feature. The simulation results show significant improvement in the performance of SVM classification algorithm by using TFCRF feature weighting method in comparative with other implemented standard feature weighting methods such as TF-based methods, IDF-based methods, TFIDF-based methods and special methods for text categorization on *inex* document collection.

Keywords

feature weighting, text categorization, SVM algorithm, feature selection, document retrieval, text mining.

[1] دانشجوی کارشناسی ارشد دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری ارتباطات، آزمایشگاه سیستم‌های هوشمند، **mmaleki@ce.aut.ac.ir**

[2] دانشیار دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فناوری ارتباطات، آزمایشگاه سیستم‌های هوشمند، **ahmad@ce.aut.ac.ir**

۱- مقدمه

طبقه بندی مستندات به مفهوم انتساب اسناد متنی بر اساس محتوی به یک یا چند طبقه از قبل تعیین شده می‌باشد؛ به بیان دیگر طبقه بندی مستندات یک فرایند یادگیری با نظارت[۱] است که درآن طبقه بندی کننده، تابع نگاشت مستندات از دامنه $D=\{d_1,d_2, \ldots, d_n\}$ به مجموعه از پیش تعریف شده طبقات $C=\{c_1, c_2, \ldots, c_m\}$ را براساس مجموعه نمونه‌های آموزشی (مستندات طبقه بندی شده) شکل می‌دهد.

در سال‌های اخیر روش‌های طبقه بندی آماری و تکنیک‌های یادگیری ماشین بسیاری از جمله درخت‌های تصمیم گیری [1]، طبقه بندی نزدیکترین همسایه(KNN)[۲] [2]، طبقه‌بندی‌کننده‌های یادگیری آماری (نظیر مدل‌های رگرسیون[3] و مدل بیزین[4])، شبکه‌های عصبی[۳] [5] ، ماشین‌های برداری پشتیبان(SVM)[۴] [6] و ... برای طبقه بندی مستندات ارائه شده اند. تعیین میزان اهمیت ویژگی یا وزن ویژگی نقش بسیار مهمی را در دستیابی به شاخص بندی با کیفیت بالا و در نتیجه دستیابی به طبقه بندی کننده خوب مستندات ایفا می‌کند [7]. در مرجع [8] نشان داده شده است که چگونگی بازنمائی مستندات به جای عملیات کرنل SVM، که یکی از بهترین روش‌های طبقه بندی مستندات به شمار می‌رود [6,9]، کارائی طبقه بندی کننده مستندات را تعیین می‌کند. این بدین معنی است که در طبقه بندی مستندات انتخاب یک روش وزن دهی مناسب ویژگی‌ها اهمیت بیشتری نسبت به انتخاب نوع طبقه بندی کننده و یا تنظیم عملیات یک طبقه کننده خاص دارد.

از جمله روش های وزن دهی ویژگی‌ها می‌توان به روش‌های مبتنی بر تعداد تکرار کلمه[۵] (TF)، روش‌های مبتنی بر تعداد تکرار کلمه در مستندات مختلف[۶] (IDF) [10]، روش‌های ترکیبی TF و IDF، روش‌های مبتنی بر الگوریتم ژنتیک و شبکه‌های عصبی [11,12]، روش‌های مبتنی بر انتخاب ویژگی[۷] [13]، روش‌های مبتنی بر مفهوم طبقه بندی اشاره کرد. اکثر این روش‌ها، روشهای استانداردی هستند که در حوزه بازیابی اطلاعات مطرح شده اند. در صورتیکه با استفاده از روش‌های وزن دهی ویژگی مخصوص طبقه بندی مستندات می‌توان به کارائی بالاتری نسبت روش‌های مذکور دست یافت.

در این مقاله یک روش جدید برای وزن دهی ویژگی خاص طبقه بندی مستندات ارائه می‌شود که در آن به ویژگی‌هایی وزن بیشتری داده می شود که بهتر توانسته باشند طبقات را از یکدیگر متمایز و تفکیک نمایند. بدین منظور در روش ارائه شده در وزن‌دهی یک ویژگی فقط به چگونگی توزیع آن در یک مستند خاص (TF) و یا چگونگی توزیع آن در مجموعه مستندات مختلف (IDF) بسنده نکرده و به چگونگی توزیع آن ویژگی در طبقات مختلف نیز توجه می‌شود.

ساختار مقاله به صورت زیر می‌باشد. در بخش ۲ برخی از روش‌های رایج وزن دهی ویژگی به طور خلاصه شرح داده می‌شوند. روش وزن دهی ویژگی پیشنهادی در بخش ۳ تشریح می‌شود. بخش ۴ شامل مشخصات محیط شبیه سازی و تحلیل نتایج حاصل از آن میباشد. بخش ۵ مقاله به ترتیب مربوط به شبیه سازی‌های انجام شده و نتیجه گیری می‌باشد.

۲- روش‌های وزن دهی ویژگی

به طور کلی می‌توان روش‌های وزن دهی متداول را به روش‌های مبتنی بر TF، روش‌های مبتنی بر IDF، روش‌های ترکیبی TFIDF و در نهایت روش‌های خاص طبقه بندی مستندات دسته بندی نمود که در این بخش خلاصه ای از نحوه عملکرد هر کدام ذکر می‌گردد.

۲-۱- روش‌های مبتنی بر TF

در این روش‌ها وزن دهی ویژگی‌ها تابعی از توزیع ویژگی‌های مختلف در هر یک از مستندات $d_i \in D$ می‌باشد.

الف- روش TF

این روش ساده و بسیار کاربردی که برای اولین بار در [14] ارائه شد که در صورت وجود ویژگی t_k در مستند d_i وزن آن برابر تعداد تکرار آن ویژگی در مستند مربوطه می‌باشد.

$$w_{ki} = tf(t_k, d_i) = \begin{cases} \#(t_k, d_i) & t_k \in \text{vector of } d_i \\ 0 & t_k \notin \text{vector of } d_i \end{cases} \quad (۲)$$

که در آن $\#(t_k, d_i)$ برابر تعداد تکرار هر ویژگی t_k در مستند d_i است.

ب- روش normTF

به طور معمول طول مستندات موجود در مجموعه D برابر نمی‌باشد، لذا در این روش برای حذف اثر طول مستند بر روی نحوه وزن دهی ویژگی‌های آن و محدود کردن مقدار وزن ویژگی‌ها بین محدوده (0,1) از نرمال سازی استفاده می‌شود.

$$w_{ki} = tf(t_k, d_i) / \sqrt{\sum_k (tf(t_k, d_i))^2} \quad (۳)$$

ج- روش logTF

در برخی مجموعه‌های داده ای متفاوت بودن ماهیت ویژگی‌ها و مقادیری که به خود اختصاص می‌دهند می‌تواند بر روی دقت و کارایی الگوریتم طبقه بندی کننده تاثیر منفی بگذارد. به همین جهت از عملگر log برای حذف این اثر نامطلوب و یکسان کردن محدوده مقادیر تخصیصی به هریک از ویژگی‌ها استفاده می‌شود.

$$w_{ki} = \log TF(t_k, d_i) = \log(tf(t_k, d_i)) \quad (۴)$$

د- روش ITF[8]

این روش برای اولین بار در [8] ارائه شد که براساس آن وزن هر ویژگی از رابطه (۵) محاسبه می‌شود

$$w_{ki} = ITF(t_k, d_i) = 1 - \frac{r}{r + tf(t_k, d_i)} \quad (۵)$$

که معمولاً مقدار r برابر ۱ قرار داده می‌شود.

ه- روش Sparck

این روش که اولین بار در [15] ارائه شد از تئوری‌های آماری برای وزن دهی به ویژگی‌ها بهره برده است.

$$w_{ki} = Sparck(t_k, d_i) = tf(t_k, d_i) * (k - \log(p_k)) \qquad (۶)$$

که در آن k تعداد کل ویژگی‌های متمایز در مجموعه D و $p_k = \sum_D tf(t_k, d_i)$ می‌باشد.

۲-۲- روش‌های مبتنی بر IDF

در این روش‌ها وزن دهی ویژگی‌ها تابعی از توزیع ویژگی t_k در داخل مجموعه مستندات D است. ایده اصلی وزن دهی در این دسته به این صورت است که هر چه تعداد مستنداتی که دارای ویژگی t_k هستند کمتر باشد، ویژگی t_k مناسبتری برای متمایز کردن مستندات از یکدیگر بوده و بایستی وزن بیشتری به خود اختصاص دهد.

الف- روش IDF سنتی

این روش که اولین بار در حوزهٔ بازیابی اطلاعات مطرح شده است [10] به شکل رابطه (۷) است.

$$w_{ki} = idf(t_k, d_i) = \log(|D| / |D(t_k)|) \qquad (۷)$$

که در آن $|D|$ تعداد کل مستندات مجموعه و $|D(t_k)|$ تعداد مستنداتی از مجموعه D می‌باشد که ویژگی t_k در آنها وجود دارد. بدیهی است که w_{ki} در رابطه (۷) با افزایش $|D(t_k)|$ کاهش می‌یابد.

۲-۳- روش‌های مبتنی بر TFIDF

این روش‌ها نیز که برای اولین بار در در حوزهٔ بازیابی اطلاعات مطرح شده سپس در طبقه بندی مستندات برای وزن دهی ویژگی‌ها از آنها استفاده شد، ح می‌باشند.

الف: روش TFIDF

روش TFIDF که از رایج ترین روش‌های وزن دهی ویژگی‌های این دسته به شمار می‌رود حاصل ترکیب روش‌های مبتنی بر TF و روش‌های مبتنی بر IDF است. به صورت زیر می‌باشد[10]:

$$w_{ki} = TFIDF(t_k, d_i) = tf(t_k, d_i) * idf(t_k, d_i) \qquad (۸)$$

ب: روش normTFIDF

برای اطمینان از اینکه همه مستندات یا طول‌های مختلف شانس برابری برای بازیابی شدن داشته باشند روش TFIDF فوق در[16] به صورت نرمال به شکل رابطه (۹) ارائه شده است

$$w_{ki} = normTFIDF(t_k, d_i) = \frac{tfidf(t_k, d_i)}{\sqrt{\sum_k (tfidf(t_k, d_i))^2}} \qquad (۹)$$

۲-۴- روش‌های مبتنی بر اطلاعات طبقات

این دسته روش‌های وزن دهی ویژگی به توزیع ویژگی t_k در مستند d_i و یا توزیع ویژگی t_k در مجموعه D بسنده نکرده و از توزیع ویژگی t_k در طبقات از پیش تعریف شده $c_j \in C$ بهره می‌برند.

الف: روش TFRF

در این روش که یکی از جدید ترین روش‌های وزن دهی ویژگی در حوزهٔ طبقه بندی مستندات می‌باشد [17] یک فاکتور ارتباط[۹] (rf) به ازای هر ویژگی t_k در طبقه c_j به صورت رابطه زیر تعریف می‌شود:

$$rf(t_k, c_j) = \log(2 + \frac{|D(t_k, c_j)|}{\sum_{m=1, m\neq j}^{|C|} |D(t_k, c_m)|}) \qquad (۱۰)$$

که در آن $c_j \in C$ ، $|D(t_k, c_j)|$ تعداد مستنداتی از مجموعه D و طبقه c_j هستند که دارای ویژگی t_k می‌باشند و $\sum_{m=1, m\neq j}^{|C|} |D(t_k, c_m)|$ مجموع تعداد مستنداتی از مجموعه D و طبقه ای غیر از طبقه c_j هستند که دارای ویژگی t_k می‌باشند. همانطور که مشخص است فاکتور rf رابطه مستقیم با تعداد مستنداتی دارد که دارای ویژگی t_k بوده و از طبقه c_j هستند و رابطه معکوس با تعداد مستنداتی دارد که دارای ویژگی t_k بوده و از طبقه ای غیر از طبقه c_j هستند. لذا وزن ویژگی t_k در مستند d_i بعد از نرمال سازی از رابطه (۱۱) محاسبه می‌شود:

$$w_{ki} = \frac{tf(t_k, d_i) * rf(t_k, c_{d_i})}{\sqrt{\sum_k (tf(t_k, d_i))^2 * (rf(t_k, c_{d_i}))^2}} \qquad (۱۱)$$

که در آن $c_{d_i} \in C$ طبقه مستند d_i است.

۳- روش وزن دهی ویژگی پیشنهادی TFCRF

اکثر روش‌های وزن دهی ویژگی فوق در ابتدا برای کاربردهای بازیابی اطلاعات مطرح شده اند و سپس در حوزه طبقه بندی مستندات به کار گرفته شده اند. لذا در این روش‌ها چگونگی توزیع ویژگی t_k در طبقه $c_j \in C$ نادیده گرفته شده است. بطورمثال همانطور که اشاره شد در روشهای مبتنی بر IDF وزن ویژگی t_k رابطه معکوسی با تعداد مستنداتی که دارای این ویژگی هستند دارد. به بیان دیگر هرچه تعداد مستنداتی که دارای ویژگی t_k هستند بیشتر باشد قدرت آن ویژگی در متمایز کردن مستندات از یکدیگر پائین تر بوده و در نتیجه وزن کمتری به آن ویژگی اختصاص داده می‌شود. اگرچه این فرض در حوزه بازیابی مستندات صحیح می‌باشد اما در حوزه طبقه بندی مستندات نیازمند اعمال اصلاحاتی است تا وزن ویژگی تابعی از طبقه مستنداتی که دارای آن ویژگی هستند نیز باشد.

واضح است هرچه تعداد مستنداتی که دارای ویژگی t_k هستند زیاد باشد ولی اکثر آن مستندات متعلق به طبقه c_j باشند، ویژگی t_k نه تنها ویژگی نامناسبی نبوده بلکه باید به عنوان یک ویژگی بسیار مناسب و مهم جهت تمایز طبقه c_j از سایر طبقات در نظر گرفته شود و وزن بالایی در آن طبقه به خود اختصاص دهد. از طرفی هرچه

مستنداتی که ویژگی t_k در آنها وجود دارد متعلق بـه طبقـاتی غیـر از طبقه c_j باشند باید وزن آن ویژگی در طبقه c_j پائین باشد.

در معیار *rf* تعریف شده در [17] راه حـل اولیـه ای بـرای مـساله فوق ارائه شده است. زیرا در آن وزن ویژگی t_k در مـستند d_i رابطـه مستقیمی بـا تعـداد مـستنداتی دارد کـه از طبقـه c_{d_i} بـوده و رابطـه معکوسی با تعداد مستنداتی دارد که از طبقه ای غیر از c_{d_i} هـستند. اما فاکتور *rf* در روش فوق مستقل از تعـداد مـستندات موجـود در هـر طبقه محاسبه می‌شود. در صورتی که توجه به همـین عامـل مـی‌توانـد کارایی طبقه بندی کننده مستندات را تا حد قابل توجهی افزایش دهد.

لذا در روش پیشنهادی ما برای وزن دهی دقیق تر به ویژگی‌ها بـه جـــای *rf* دو فـــاکتور *positiveRF* (فـــاکتور ارتبـــاط مثبـــت) و *negativeRF* (فاکتور ارتباط منفـی) تعریـف مـی‌شـود. *positiveRF* نسبت تعداد مستنداتی از طبقه c_j را که ویژگی t_k را دارنـد بـه کـل مستندات آن طبقه نشان می‌دهد و *negativeRF* نسبت مجموع تعداد مستنداتی از طبقه غیر c_j را که ویژگی t_k را دارند بـه کـل مجمـوع مستندات طبقات غیر c_j را نشان می‌دهد که بـه صـورت زیـر تعریـف می‌شوند:

$$positiveRF(t_k, c_i) = |D(t_k, c_j)| / |D(c_j)| \qquad (۱۲)$$

$$negativeRF(t_k, c_i) = \frac{\sum_{m=1, m\neq j}^{|C|} |D(t_k, c_m)|}{\sum_{m=1, m\neq j}^{|C|} |D(c_m)|} \qquad (۱۳)$$

کـه در روابـط فـوق $|D(c_j)|$ تعـداد مـستندات طبقـه c_j و $|D(t_k, c_j)|$ تعداد مستنداتی از مجموعـه D و طبقـه c_j کـه دارای ویژگی t_k می‌باشند است.

از دو رابطه (۱۲) و (۱۳) مقـدار ارزش فـاکتور ارتبـاط هـر طبقـه (*crfValue*[10]) به طور کلی به صورت زیر تعریف می‌شود

$$crfValue(t_k, c_i) = \frac{positiveRF(t_k, cj)}{negativeRF(t_k, c_j)} \qquad (۱۴)$$

مشخص است ارزش فاکتور ارتباط هـر طبقـه رابطـه مـستقیم بـا فاکتور ارتباط مثبت و رابطه معکوس با فاکتور ارتباط منفی دارد.

رابطه پیشنهادی برای وزن دهـی ویژگـی t_k در مـستند d_i بـه صورت رابطه (۱۵) است

$$w_{ki} = \log(tf(t_k, d_i) * crfValue(t_k, c_{d_i})) \qquad (۱۵)$$

که در آن $c_{d_i} \in C$ طبقه مستند d_i است.

برای از بین بردن اثر طول مستند بر دقت و کـارایی طبقـه بنـدی کننده، از نرمال کردن استفاده شده تا وزن ویژگـی‌هـا در دامنـه (0,1) محدود شود. درنتیجه رابطه نهایی پیـشنهادی بـصورت رابطـه (۱۶) در خواهد آمد:

$$w_{ki} = \frac{\log(tf(t_k, d_i) * crfValue(t_k, c_{d_i}))}{\sqrt{\sum_k (\log(tf(t_k, d_i) * crfValue(t_k, c_{d_i})))^2}} \qquad (۱۶)$$

نشان داده خواهد شد که این روش وزن دهی ویژگی بـرای طبقـه بندی مستندات نسبت به سایر روش‌های شرح داده شده در بخش ۲ از کارایی بالاتری برخوردار است.

۴- شبیه سازی و نتایج

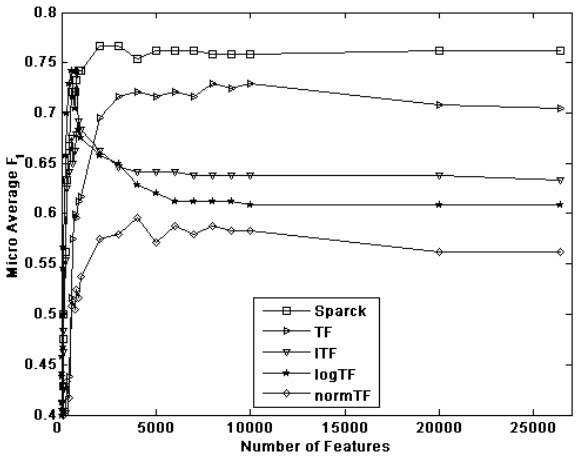
در محیط شبیه سازی طراحی شده، طبقه بنـدی مجموعـه مـستندات *inex* (شـامل ۱۲۱۰۷ مقالـه علمـی از ۱۸ مجلـه انتـشارات انجمـن کامپیوتر IEEE از سـال ۱۹۹۵ تـا ۲۰۰۲ بـه فرمـت XML [18])بـه روش SVM به ازای کلیه روش‌های وزن دهی ویژگی شـرح داده شـده در بخش ۲ و روش پیشنهادی TFCRF پیاده سازی شده و در نهایـت بر اساس معیار ارزیابی میانگین میکرو F_1 با یکدیگر مقایسه شده اند.

از آنجا که در این مقاله هدف مقایسه روش‌های مختلف وزن دهی ویژگـــی در حـــوزه طبقـــه بنـــدی مـــستندات اســـت لـــذا ۶ طبقه(tc,td,tg,tk,tp,ts) مجموعه *inex*، و از هر طبقه ۱۲۰ مستند بـه صورت تصادفی به عنوان مجموعـه مـستندات D انتخـاب شـده انـد. تقسیم بندی مستندات آموزشی و آزمایشی به نسبت ۲/۳ و ۱/۳ یعنی ۴۸۰ مستند آموزشی (TrD) و ۲۴۰ مستند آزمایشی (TeD) در کل مجموعه D است. تعداد کل ویژگی‌های متمـایز مجموعـه مـستندات آموزشی بعد از حذف stopwordها، اعداد و علائم نگارشی از مجموعـه ویژگـی‌هـا و اعمـال Porter Stemmer بـرای ریـشه یـابی ویژگـی‌هـا ۲۶۴۳۴ ویژگی است.

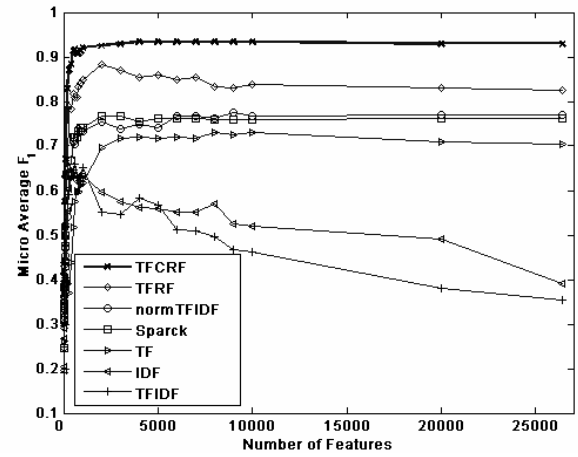
به علت تنوع روش‌های وزن دهی ویژگـی، ابتـدا روش‌هـای وزن دهـی مبتنی بر TF با یکدیگر مقایسه شـده و در نهایـت دو روش بـا کـارائی بـالاتر از بـین آنهـا بـا سـایر روش‌هـای وزن دهـی و روش پیـشنهادی TFCRF مقایسه می‌گردند.

در شکل (۱) میانگین میکرو F_1 مربوط بـه روش‌هـای وزن دهـی مبتنی بر TF نشان داده شده است. همانگونه کـه مـشاهده مـی‌گـردد روش Sparck و پس از آن روش TF در مقایسه با سایر روش‌ها کارائی بهتری از خود نشان می‌دهند. بهترین مقدار میـانگین F_1 در روش TF، ۰/۷۳ بـه ازای ۸۰۰۰ ویژگـی اسـت در حالیکـه ایـن مقـادیر در روش Sparck به ازای ۲۰۰۰ ویژگـی ۰/۷۶۷ مـی‌باشـد. روش logTF علـی رغم انتظار به دلیل یکسان بودن ماهیت ویژگی‌هـا (کلمـات موجـود در مستندات) و مقادیر انتسابی به آنها (تعداد تکرار در مستندات) در حوزهٔ طبقه بندی مستندات نسبت به TF کارائی پـائین تـری از خـود نـشان داده است. دلیل پـائین بـودن کـارائی normTF در مقایـسه بـا TF را مـی‌تـوان در یکـسان بـودن تقریبـی طـول مـستندات مجموعـه *inex* (مقالات IEEE) دانست.

در شکل (۲) میانگین میکرو F_1 روش Sparck و TF از روش‌های وزن دهی ویژگی مبتنی بر TF، روش‌های مبتنی بر IDF و TFIDF، روش TFRF و روش TFCRF پیشنهادی نشان داده شده است.



شکل (۱): مقایسه کارائی روش‌های وزن دهی ویژگی مبتنی بر TF



شکل (۲): مقایسه کارائی وزن دهی ویژگی پیشنهادی با روش‌های موجود

همانطور که مشاهده می‌شود روش IDF و روش TFIDF در مقایسه با سایر روش‌ها کارائی پائین تری از خود نشان می‌دهند. دلیل این امر را می‌توان در بکارگیری غلط فاکتور idf در بحث طبقه بندی مستندات دانست زیرا همانطور که در بخش ۲-۲ اشاره شد در این روش‌ها وزن هر ویژگی رابطه معکوس با تعداد مستنداتی دارد که دارای آن ویژگی هستند و این مفهوم در حوزه طبقه بندی مستندات نیاز به اصلاحاتی دارد. روش‌های وزن دهی TF و Sparck نسبت به روش‌های IDF و TFIDF کارائی بهتری از خود نشان می‌دهند اما از آنجا که در این روش‌ها وزن دهی ویژگی‌ها فقط تابعی از توزیع آنها در مستندات مختلف است نسبت به روش TFRF و TFCRF کارائی پائینتری دارند. روش TFRF به دلیل آنکه در وزن دهی ویژگی‌ها علاوه بر توجه به چگونگی توزیع آنها در مستندات مختلف به چگونگی توزیع آنها در طبقات مختلف نیز توجه دارد در مقایسه با سایر روش‌های وزن دهی ویژگی (بجز روش TFRCF) کارائی بالاتری برخوردار است. مشخص است که روش پیشنهادی TFCRF با ازای تعداد ویژگی‌های مختلف کارائی بسیار بالاتری نسبت به سایر روش‌ها از خود نشان می‌دهد. بهترین مقدار میانگین میکرو F_1 مربوط به این روش است که به ازای ۴۰۰۰ ویژگی ۰/ ۹۳۳ می‌باشد در حالیکه این مقادیر در روش TFRF به ازای ۲۰۰۰ ویژگی ۰/ ۸۸۳ می‌باشد. این نتیجه، تحلیل ارائه شده در بخش ۳ را که محاسبه مقدار ارزش فاکتور ارتباط طبقه (*crfValue*) می‌تواند قدرت ویژگی‌ها را در متمایز کردن طبقات از یکدیگر بیشتر کند تایید می‌نماید.

۵- نتیجه گیری

در این مقاله یک روش جدید برای وزن دهی ویژگی خاص حوزه طبقه‌بندی مستندات به نام TFCRF ارائه شد که در آن برای وزن دهی ویژگی‌ها علاوه بر توجه به چگونگی توزیع آنها در مستندات مختلف و در کل مستندات مجموعه به چگونگی توزیع آنها در طبقات مختلف نیز توجه می‌شود. همچنین در روش پیشنهادی برای محاسبه مقدار ارزش فاکتور ارتباط هر طبقه (*crfValue*) به تعداد مستندات موجود در هر طبقه به عنوان عامل مهمی در وزن‌دهی ویژگی توجه شده است. در ادامه عملکرد و کارائی روش پیشنهادی با روش‌های رایج وزن دهی ویژگی نظیر روش‌های مبتنی برTF، روش‌های مبتنی بر IDF، روش‌های ترکیبی TFIDF و روش‌های خاص طبقه بندی مقایسه گردید. نتایج شبیه سازی نشان دهنده بهبود قابل توجهی در کارایی الگوریتم طبقه بندی کننده SVM با بکارگیری روش وزن دهی ویژگی ارائه شده جدید TFCRF در مقایسه با سایر روش‌های وزن دهی ویژگی بر روی مجموعه مستندات *inex* است.

مراجع

[1]Apte, C., Damerau, F., Weiss, S., "*Text Mining with Decision Rules and Decision Trees*", The Conference on Automated Learning and Discovery, Workshop 6:Learning from Text and the Web, 1998.

[2]Creecy, R.M. et al., "*Trading MIPS and Memory for Knowledge Engineering: Classifying Census Returns on the Connection Machine*", Communications of the ACM, Vol. 35, No. 8, pp. 48–63, 1992.

[3]Yang, Y., Chute, C.G., "*An Example-Based Mapping Method for Text Categorization and Retrieval*", ACM Transaction on Information Systems (TOIS), Vol. 12, No. 3, pp. 252-277, 1994.

[4]Koller, D., Sahami, M., "*Hierarchically classifying documents using very few words*", In the 14th International Conference on Machine Learning (ICML97), pp. 170–178, Nashville, US, 1997.

[5]Wiener, E.D., *A Neural Network Approach to Topic Spotting in Text*, Master's thesis, Department of Computer Science, University of Colorado at Boulder, US, 1995.

[6]Joachims, T., "*Text Categorization with Support Vector Machines: Learning with Many Relevant Features*", In European Conference on Machine Learning (ECML), 1998.

[7]Zhang, J. and Nguyen, T.N., "*A New Term Significant Weighting Approach*", Journal of Intelligent information system, Vol. 24. No. 1, pp. 61-85, 2005.

[8]Leopold, E. and Kindermann, J., "*Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?*", Machine Learning, Vol. 46, No. 1-3, pp. 423-444, 2002.

[9]Yang, Y. and Liu, X., "*A Re-Examination of Text Categorization Methods*", The 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49. ACM Press, 1999.
[10]Salton, G., Yang, C.S., "On the Specification of Term Values in Automatic Indexing", Journal of Documentation, Vol. 29, No. 4, pp. 351-357, 1973.
[11]Robertson, A.M. and Willett, P., "*An Upperbound to the Performance of Ranked-Output Searching: Optimal Weighting of Query Terms Using a Genetic Algorithm*", Journal of Documentation, Vol. 52, pp. 405–420, 1996.
[12]Boger, Z., Kuflik, T., and Shoval, P., "*Automatic Keyword Identification by Artificial Neural Networks Compared to Manual Identification by Users of Filtering Systems*", Info. Proc. and Management, Vol. 37, No. 2, pp. 187–198, 2001.
[13]Debole, F., and Sebastiani, F., "*Supervised term weighting for automated text categorization*", the 2003 ACM symp. on Applied computing, pp. 784-788. ACM Press, 2003.
[14]Luhn, H.P, "A Statistical Approach to the Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No. 4, pp. 309–317, 1957.
[15]Sparck Jones, K. "*Indexing Term Weighting*", Information Storage and Retrieval, Vol. 9, pp. 619–633, 1973.
[16]Salton, G., Allan, J., and Singhal, A., "*Automatic Text Decomposition and Structuring*", Information Processing and Management, Vol. 32, No. 2, pp. 127–138, 1996.
[17]Lan, M., Sung, S.Y., Low, H.B., Tan, C.L., "*A Comparative Study on Term Weighting Schemes for Text Categorization*", IEEE International Conference on Neural Networks (IJCNN05), pp. 546-551, 2005.
[18]INitiative for the Evaluation of XML Retrieval (INEX), http://*inex*.is.informatik.uni-duisburg.de

زیرنویس‌ها

[1] Supervised Learning
[2] K-Nearest Neighbor
[3] Neural Networks
[4] Support Vector Machine
[5] Term Frequency
[6] Inverse Document Frequency
[7] Feature election
[8] Inverse Term Frequency
[9] Relevancy Factor
[10] Category Relevancy Factor Value