فصلنامه
نظام ها و خدمات اطلاعاتی
سال اول، شماره ۴، پائیز ۱۳۹۱ص ص ۹۵-۱۰۴.

وزن‌دهی به اصطلاحات و نقش آن در بازیابی اطلاعات

زهرا کاظم پور *
عضو هیأت علمی دانشگاه پیام نور و دانشجوی دکتری تخصصی علم اطلاعات و دانش شناسی دانشگاه تهران

دکتر فاطمه فهیم نیا
استادیار گروه علم اطلاعات و دانش شناسی دانشگاه تهران

تاریخ دریافت: ۱۳۹۱/۷/۳۰ **تاریخ پذیرش:** ۱۳۹۱/۹/۱۵

چکیده

هدف: امروزه رشد و انفجار اطلاعات منجر به افزایش نیاز به ایجاد روش‌های کارآمد در بازیابی اطلاعات شده است. بنابراین، به منظور افزایش عملکرد بازیابی اطلاعات، شیوه‌های جدیدی ایجاد شده است. یکی از مهم‌ترین این شیوه‌ها وزن‌دهی به اصطلاحات مدارک موجود در پایگاه‌ها یا مخازن اطلاعات است. از این شیوه به منظور رتبه‌بندی مدارک استفاده می‌شود. هدف از مطالعه حاضر، شناسایی نقش و اهمیت وزن‌دهی در بهبود بازیابی اطلاعات است.

روش: گردآوری اطلاعات در این پژوهش با استفاده از متون تخصصی انجام شده است. بنابراین روش پژوهش، کتابخانه‌ای است.

یافته‌ها: طرح‌های پیچیده وزن‌دهی نسبت به شیوه‌های ساده که صرفاً بر اساس حضور هر اصطلاح در متن می‌باشد، از عملکرد بهتری در بازیابی اطلاعات برخوردارند. بنابراین پژوهشگران در وزن‌دهی به اصطلاح‌ها، در جستجوی شیوه‌هایی با کارایی بالاتر می‌باشند که در عمل موجب بهبود بازیابی مدارک می‌شوند.

نتیجه گیری: در مجموع وزن‌دهی اصطلاحات عملکرد بازیابی اطلاعات را ارتقا می‌بخشد. با این وجود، وزن‌دهی به اصطلاح تنها راه افزایش و بهبود کارکرد سیستم‌های بازیابی اطلاعات نیست و مجموعه‌ای از عوامل در بهبود این سیستم‌ها تأثیرگذار هستند.

کلیدواژه‌ها: وزن‌دهی، اصطلاحات، بازیابی اطلاعات.

* نویسنده رابط: zahrakazempour@ut.ac.ir

مقدمه

نظام‌های اطلاعاتی در طول تاریخ پیدایش خود به دنبال این هدف بودند که میزان انطباق خواسته کاربران و مدارک ارائه‌شده به عنوان نتایج جستجو را افزایش دهند. این کار به طور معمول از طریق تطابق کلیدواژه‌های عبارت جستجوی کاربران و کلیدواژه-های نماینده مدارک موجود در مخزن مدارک یا پایگاه داده‌های مراکز اطلاع‌رسانی انجام می‌شود. از آنجا که خواسته‌های واقعی کاربران به طور کامل در قالب زبان مصنوعی ماشین و همچنین کلیدواژه‌های منفرد و حتی عبارت‌های جستجو تبدیل نمی‌شود، در نتیجه، طراحان نظام‌های اطلاعاتی و متخصصان حوزه اطلاعات و اطلاع‌رسانی همیشه به دنبال شیوه‌های بهینه برای افزایش تطابق پرسش کاربران با مدارک موجود بوده‌اند (حسن زاده، ۱۳۸۳).

بدین ترتیب، در تمام سیستم‌های بازیابی اطّلاعات، چنانچه مدرک بازیابی شده در قضاوت کاربر، مورد توجه وی واقع گردد، آن مدرک به عنوان مدرک مربوط و در غیر این‌صورت، مدرک نامربوط شناخته می‌شود. از آنجا که عوامل بسیاری، قضاوت درباره ربط را با استفاده از روش‌های پیچیده تعیین می‌کنند، یک سیستم بازیابی اطّلاعات نمی‌تواند به طور دقیق تمام مدارک مربوط را انتخاب نماید. بنابراین، سیستم باید روش‌هایی را بپذیرد که رتبه‌بندی مدارک را به ترتیب احتمال استفادهٔ کاربر از آنها آسان کند.

در روش‌های معمول یک اصطلاح می‌تواند هم به یک مدرک اختصاص یابد و هم می‌تواند اختصاص نیابد. اگر چه این مسئله فرایند نمایه‌سازی را تسهیل می‌کند، اما برای کاربر پایگاه اطلاعاتی مشکلاتی را بوجود می‌آورد؛ کاربر نمی‌تواند راهبرد کاوش موفقی را طراحی کند که از طریق آن بتواند مدارکی که در آنها یک موضوع با قوت بحث‌شده را از مدارکی که در آنها موضوع بطور فرعی مورد بحث قرار گرفته است، تشخیص دهد (لنکستر، ۱۳۸۲، ص. ۲۶۵). یکی از روش‌های مناسب برای حل این مشکل، محاسبهٔ همبستگی اصطلاحات، بر اساس وزن‌دهی اصطلاحات است. در یک سیستم بازیابی اطّلاعات، معمول است که یک مدرک به وسیلهٔ کلیدواژه‌ها یا واژه‌های موضوعی نمایانده شود. کلیدواژه‌ها معمولاً در فرایند نمایه‌سازی، از متن یا چکیدهٔ مدرک استخراج می شوند. علاوه بر گزینش اصطلاحات برای بازنمون مدارک، معمولاً به هر اصطلاح وزنی می‌دهند تا اهمیّت آن اصطلاح خاص را در مدرک نشان دهد (مهراد، کلینی، ۱۳۸۶). با توجه به اینکه در شیوه معمول نمایه سازی معمولاً بین اصطلاحات ارتباطات صریح و روشنی وجود ندارد و اینکه نمایه ساز مشخص نمی‌کند که چرا برخی واژه ها نسبت به دیگر واژه ها اهمیت بیشتری دارند، لذا، توجه به وزن‌دهی اصطلاحات ضرورت می‌یابد. بنابراین در ادامه مباحثی در ارتباط با وزن‌دهی به اصطلاحات ارائه می‌شود.

تعریف وزن‌دهی به اصطلاحات

وزن‌دهی به اصطلاحات نقش مهمی در عملکرد سیستم بازیابی اطلاعات ایفا می‌کند. وزن یک اصطلاح، اطلاعاتی در مورد ربط آن اصطلاح به مدرک فراهم می‌کند (Bisht, Srivastava, Dhami, 2010). از آنجا که اصطلاحات متفاوت در متن دارای اهمیت مختلفی هستند، از یک نشانگر مهم یعنی «وزن اصطلاح» استفاده می‌شود که همراه هر اصطلاح است. این وزن با روش وزن‌دهی تعیین می‌گردد. وزن‌دهی روشی است که در آن چگونگی توزیع اصطلاحات و تکرار آنها در مدرک بررسی می‌شود و برای اصطلاحات وزن عددی تولید می‌کند. این وزن به معنای درجه اهمیت و اعتبار یک واژه در مدارک مختلف

می‌باشد. به عبارت دیگر، این اوزان بر فراوانی رخداد اصطلاح در کل مجموعه مدارک و تعداد حضور یک اصطلاح در مدرک خاص مبتنی است. با افزایش فراوانی اصطلاح در یک مدرک، وزن آن افزایش می‌یابد و برعکس زمانی که فراوانی اصطلاح در مجموعه مدارک بیشتر باشد، این وزن کاهش می‌یابد (مهراد، کلینی، ۱۳۸۶). در تعریفی دیگر، وزن‌دهی معادل ارزش گذاری اصطلاحات موجود در عبارت جستجو به کار رفته است که طی آن بالاترین ارزش و وزن به اصطلاحی داده می‌شود که مربوط ترین یا مفیدترین اصطلاح برای درخواست جستجو باشد (کوکبی، ۱۳۸۷). بنابراین وزن‌ها اهمیت نسبی موضوعات مختلف را در مدارک نشان می‌دهند. به عبارتی، با وزن‌دهی، به هر مفهوم نشانه ای از اهمیت نسبی آن در مدرک داده می‌شود (Radecki, 1983).

هدف و فایده وزن‌دهی به اصطلاحات

هدف وزن‌دهی، پدید آوردن امکان تطبیق خودکار پرسش‌ها با رکوردهای مدارک از طریق محاسبه ارزش تطبیق ریاضی است (کوکبی، ۱۳۸۷). به عبارتی وزن‌های محاسبه شده برای تعیین میزان مشابهت بین عبارت پرسش و مدارک، مورد استفاده قرار می‌گیرد (حسن زاده، ۱۳۸۳). بر این اساس می‌توان گفت هدف وزن‌دهی، بهبود بازیابی اطلاعات و افزایش میزان ربط می‌باشد.

همچنین وزن‌دهی موجب می‌شود کاربر بتواند از میزان ارتباط مدارک بازیابی شده با پرسش خود آگاهی یابد، زیرا بر اثر وزن‌دهی، مدارک بازیابی شده بر اساس درجه مشابهت بین عبارت پرسش و مدارک، مورد استفاده قرار می‌گیرند.

- وزن‌دهی موجب سادگی تشخیص میزان تشابه میان عبارت پرسش و مدارک موجود در مجموعه می‌شود.
- عملکرد بازیابی را ارتقا دهد.
- منجر به رتبه بندی مدارک بازیابی شده می‌شود و در نتیجه کاربر می‌تواند در مورد نتایج ارائه شده قضاوت کند.
- در ارتقای ربط کاربرد دارد (حسن زاده، ۱۳۸۳).
- به کاربر امکان می‌دهد که اهمیت هر اصطلاح را نسبت به اصطلاح دیگر معین کند (میدو و دیگران، ۱۳۹۰، ص. ۳۷۸).

عوامل مؤثر بر وزن‌دهی اصطلاحات

عوامل متفاوتی بر میزان وزن اصطلاح مؤثر می‌باشد که عبارتند از (Salton, Buckley, 1988):

۱. تعداد تکرار اصطلاح در یک مدرک؛ هر چه این تعداد بیشتر باشد اهمیت اصطلاح بیشتر است.
۲. تعداد مدارکی که اصطلاح در آنها ظاهر شده است؛ هر چه این تعداد بیشتر باشد اهمیت اصطلاح در تمایز مدارک کمتر خواهد بود.
۳. محل ظاهرشدن اصطلاح؛ معمولاً اصطلاحی که در عنوان یک مدرک باشد مهم‌تر از اصطلاح‌های درون متن است. به ترتیب، اهمیت بیشتر برای اصطلاح‌هایی است که در عنوان، چکیده، کلیدواژه‌ها و سرفصل‌های متن ظاهر شده‌اند.
۴. نمادهای ظاهری اصطلاح (نمادهایی مانند فونت و قلم نوشتاری) ؛ معمولاً اصطلاح‌های نوشته شده با قلم خاص مهم‌تر از اصطلاح‌هایی است که بدون این ویژگی باشند.
۵. طول مدرک؛ هر چه طول مدرک بیشتر باشد، بسامد اصطلاحات در آن افزایش می‌یابد و عدد وزن بیشتر می‌شود.

شیوه‌های رایج وزن‌دهی به اصطلاحات

وزن‌دهی به اصطلاحات گامی مهم در موفقیت بازیابی اطلاعات است. در دهه‌های گذشته با تقاضاهایی که برای جستجو در وب و پایگاه‌های اطلاعاتی وجود داشته است، علاقه فزاینده‌ای به این حوزه ایجاد شده است. در دهه‌های گذشته شیوه‌هایی گوناگونی از وزن‌دهی ایجاد و کارایی آنها آزمایش شده است. اغلب این شیوه‌ها آماری می‌باشند و بر این فرضیه استوار هستند که رفتار آماری یک اصطلاح در مدارک مجزا یا مجموعه مناسبی از مدارک، توانایی اصطلاح را برای نمایش محتوای مدرک یا متمایز کردن آن از مدارک دیگر نشان می‌دهد. در ادامه تعدادی از رایج ترین شیوه‌های وزن‌دهی به اصطلاحات معرفی می‌شود.

ساده ترین شیوه وزن‌دهی، شیوه دودویی می‌باشد که بدین صورت نشان داده می‌شود:

وزن اصطلاح
- ۱ وقتی بسامد اصطلاح در مدرک بزرگتر از صفر است. F>0
- 0 وقتی بسامد اصطلاح در مدرک برابر با صفر است. F=0

این شیوه نمی‌تواند توصیفگر مناسبی برای یک واژه باشد. به عنوان مثال هنگامی که اصطلاح الف در یک مدرک ۱۵ بار ظاهر شود و اصطلاح ب در آن مدرک یک بار ظاهر شود، وزن هر دو اصطلاح برابر با یک است. بنابراین این شیوه نمی‌تواند اهمیت اصطلاحات را به درستی نشان دهد. از این رو شیوه‌های دیگری از وزن‌دهی ایجاد شد (Bisht, Srivastava, Dhami, 2010).

روش «بسامد اصطلاح» که به روش TF[1] معروف است برای نخستین بار در سال ۱۹۶۸ توسط سالتون[2] مطرح شد (Jones, 2004). این روش تعداد دفعاتی که یک اصطلاح در یک مدرک ظاهر می‌شود را نشان می‌دهد. به عبارتی نشان می‌دهد که وزن یک اصطلاح معین به بسامد این اصطلاح در یک مدرک معین وابسته است (Ruch, Baud, Geissbuhler, 2002). بنابراین بسامد بالای حضور یک اصطلاح در مدرکی خاص نشان دهنده اهمیت بالای آن اصطلاح در مدرک است (Raju, Sukavasi, Chava, 2011).

شیوه محاسبه این روش بر اساس فرمول ذیل می‌باشد (Cheng, 2006):

$$w_{ij} = tf_{ij}$$

w_{ij} به معنای وزن اصطلاح است.

tf_{ij} به معنای تعداد حضور اصطلاح i در مدرک j است.

در سال ۱۹۷۲ کارن اسپارک جونز[3] مقاله ای منتشر کرد که در آن شیوه جدیدی از وزن‌دهی با عنوان «معکوس بسامد مدرک» IDF[4] پیشنهاد شده بود (Robertson, 2004) . IDF یکی از شیوه‌های رایج وزن‌دهی و بر اساس این ایده است که چنانچه یک اصطلاح در تعداد کمی از مدارک موجود در مجموعه ظاهر شود، احتمالاً یک تفکیک کننده مناسب برای این مدارک است. فرمول آن عبارت است از (Nanas, Uren, Roeck, 2003):

$$IDF = \frac{\text{تعداد کل مدارک}}{\text{تعداد مدارکی که اصطلاح در آنها وجود دارد}}.$$

1. Term Frequency
2. Salton
3. Karen Sparck Jones
4. Inverse Document Frequency

یا به عبارتی: $$IDF = \log\frac{N}{n}$$

از لگاریتم برای کوچک کردن وزن استفاده می‌شود.

در این شیوه، اصطلاحات پر بسامد در مجموعه مدارک دارای وزن کمتر می باشند. به عبارتی، هر چه فراوانی اصطلاحی در کل پایگاه اطلاعاتی بیشتر باشد، میان مدارک مربوط و سایر مدارک تفاوت کمتری می‌گذارد. بر این اساس اصطلاحی که به ندرت به کار می رود، وزن بیشتری را به خود اختصاص می‌دهد (پولیت، ۱۳۸۳، ص. ۱۱۵). اما این شیوه را نمی توان به عنوان تنها معیار وزن‌دهی و به تنهایی به کار برد؛ زیرا این شیوه نمی‌تواند به رکوردهایی که دارای عباراتی یکسان هستند، وزن‌های مختلف دهد و در نتیجه آنها را رتبه‌بندی کند (لارج، تد، هارتلی، ۱۳۸۲، ص. ۲۰۳). از این رو در سال ۱۹۷۳ شیوه ترکیبی TF×IDF توسط سالتون و یانگ[1] پیشنهاد شد (Robertson, 2004) . استفاده از این شیوه ترکیبی موجب عدم تأثیر طول مدرک بر نتیجه بازیابی می‌شود. فرمول این شیوه عبارت است از (Cheng, 2006):

$$w_{ij} = tf_{ij} \times \log\frac{N}{n}$$

tf_{ij} تعداد ظهور اصطلاح i در مدرک j است.

N تعداد کل مدارک است.

n تعداد مدارکی است که شامل اصطلاح i هستند.

امروزه استفاده از این روش نسبت به سایر روش‌ها رواج بیشتری دارد.

نقطه ضعف روش وزن‌دهی TF

یکی از مشکلاتی که در استفاده از روش وزن دهی TF وجود دارد این است که بسامد یک اصطلاح در متن‌های طولانی، زیاد و در متن‌های کوتاه کم می‌باشد که موجب می‌شود مدارک طولانی بیشتر از مدارک کوتاه بازیابی شوند (گراسمن، فریدر، ۱۳۸۴، ص ۲۷). بدین منظور و برای حل این مسئله عاملی به نام عامل نرمال سازی در فرمول‌های وزن‌دهی مورد استفاده قرار می‌گیرد که موجب مساوی شدن طول مدارک می‌شود. (Salton, Buckley, 1988).

به عنوان مثال فرمول نرمال شده ی TF به صورت زیر می‌باشد: $$Normalized\ \ TF_i = \frac{TF_i}{\sqrt{\sum TF_j{}^2}}$$

فرمول نرمال شده ی TF×IDF نیز عبارت است از: $$Normalized\ \ TF_i \times IDF_i = \frac{tf_i \times \log\frac{N}{n_i}}{\sqrt{\sum (tf_j \times \log\frac{N}{n_j})^2}}$$

که در آن:

N: تعداد کل مدرک در مجموعه است.

n_i: تعداد مدارکی در مجموعه است که دارای اصطلاح i می باشند.

n_j: تعداد اصطلاح‌های نمایه ای مجزا در متن است.

این فرمول در وزن‌دهی نرمال بسیار مؤثر می‌باشد (Kang, Kim, Lee, 2005).

1. Yang

نقطه ضعف روش وزن‌دهی IDF

همانگونه که اشاره شد در روش وزن‌دهی IDF باید اطلاعاتی در مورد کل مجموعه ی مدارک داشته باشیم. در استفاده از این روش تعداد مدرک موجود عاملی مهم در محاسبه وزن می‌باشد. مشکلی که در اینجا رخ میدهد این است که تعداد مدارک درون یک مجموعه دائماً در حال تغییر است زیرا مدارک جدید به مجموعه اضافه و مدارک قدیمی از آن حذف می‌شوند (Witschel, 2008) .

پژوهش‌ها در این زمینه نشان داده است حداقل تا زمانی که مدارک جدید اصطلاحات جدید بسیار زیاد نداشته باشند، اضافه شدن آنها به مجموعه موجب کاهش کارایی فرمول وزن‌دهی نمی‌شود. اما در موارد دیگر روزآمد نبودن مجموعه در محاسبه اثر منفی می‌گذارد. از این رو راهکارهایی پیشنهاد شده است که عبارتند از (Kowalski, 1998):

۱. تغییرات نادیده گرفته شوند و وزن‌ها بر اساس ارزش‌های جدید محاسبه شوند؛ بدین معنا که برای مجموعه ای که روزآمد شده است به صورت دوره ای وزن‌دهی مجدداً انجام گیرد.
۲. از یک مقدار ثابت استفاده کنیم و هنگامی که تغییرات به آستانه ی معینی رسید از مقدار جدید استفاده کنیم. به عبارتی از آنجا که تغییرات کوچک در مقادیر، تأثیر کمی بر وزن نهایی دارد؛ پس وزن‌دهی مجدد زمانی انجام می‌گیرد که تغییرات به میزان کافی روی داده باشد و مقدار قبلی بر وزن نهایی تأثیر داشته باشد.
۳. مقدار متغیرهای بدون تغییر مانند بسامد واژه در یک مدرک را حفظ کرده و در هنگام جستجو آخرین وزن‌ها را برای یک مدرک محاسبه کنیم؛ به عبارتی هر بار که یک پرسش وارد سیستم می‌شود بخش TF که مقداری ثابت است را باید از قبل داشته باشیم و تنها بخش IDF را برای آن محاسبه کنیم تا سریع تر وزن نهایی را به دست آوریم.

علاوه بر راهکارهای ذکر شده، در حال حاضر محققان در حال انجام پژوهش‌ها و ابداع روش‌های دیگری هستند که در آنها وزن‌دهی به تعداد کل مدارک در مجموعه وابستگی کمتری داشته باشند تا از بروز مشکل در محاسبه ی وزن جلوگیری شود.

سایر روش‌های وزن‌دهی

سال‌ها بعد از ابداع روش‌های وزن‌دهی TF، IDF و TF×IDF، به علت کاستی‌هایی که این فرمول‌ها در موارد خاص در محاسبه وزن اصطلاح‌ها داشتند، محققان به ابداع روش‌های دیگری پرداختند. این روش‌ها بسیار متنوع هستند و هر کدام در شرایط ویژه ی خود مورد استفاده قرار می‌گیرند. اغلب آنها بر پایه فرمول‌های اولیه ذکر شده می باشند. در ادامه به صورت مختصر به دو روش دیگر اشاره می‌کنیم.

روش وزن‌دهی به ساختار

مدارک دارای ساختار هستند که سیستم‌های بازیابی اطلاعات معمولاً آن را نادیده گرفته اند. از آنجا که بخش‌هایی از مدرک نسبت به سایر قسمت‌ها از اهمیت بیشتری برخوردارند؛ بنابراین ساختار مدرک باید در رتبه بندی آن مورد توجه قرار گیرد. به عنوان مثال اطلاعاتی که در چکیده مدرک است نسبت به بخش‌های دیگر مهم‌تر می‌باشد. پژوهشگران این نکته را مورد توجه قرار دادند که اصطلاح‌ها بر اساس مکان ظهور باید وزن متفاوتی بگیرد. بر این اساس روش وزن‌دهی به ساختار اولین بار در سال ۱۹۹۳

پیشنهاد شد. در عمل این شیوه ی وزن‌دهی با استفاده از الگوریتم‌های پیچیده ی ژنتیک انجام گرفت که در آن برای درک شباهت‌ها، هر ساختار مدرک در یک کروموزوم در یک الگوریتم نشان داده شده است. وزن‌دهی در این روش که با استفاده از فرمول‌های پیچیده انجام می‌گیرد ساختار مدرک را مورد توجه قرار می‌دهد (Trotman, 2005).

روش وزن‌دهی بر اساس بافت متن[1]

در این روش عواملی مانند نقش معنایی یک اصطلاح، فاصله میان دو اصطلاح و محل حضور اصطلاح با استفاده از فرمول‌های پیچیده محاسبه می‌شود. سپس عدد به دست آمده از آن با عدد حاصل از فرمول TF×IDF جمع بسته می‌شود و وزن نهایی محاسبه می‌شود (Ernandes, et.al, 2007).

فرایند وزن‌دهی خودکار به اصطلاحات

همانطور که اشاره شد روش ترکیبی TF×IDF در وزن‌دهی به اصطلاحات، کاربرد بیشتری دارد. بر این اساس در طراحی فرایند خودکار وزن‌دهی معمولاً این شیوه را مورد استفاده قرار می دهند. مراحل وزن‌دهی خودکار به اصطلاحات به ترتیب عبارتند از (Raju, Sukavasi, Chava, 2011; Hong and et. al., 2005; Drucker, Shahrary, Gibbon, 2002):

۱ – انتخاب یک مدرک از پایگاه اطلاعاتی به منظور وزن‌دهی؛

۲ – حذف واژه‌های بدون ارزش معنایی؛ در این مرحله با استفاده از یک «فهرست بازدارنده»[2] که قبلاً تهیه شده است، واژه‌های غیر مهم مانند حروف اضافه، افعال کمکی، کلمات یک یا دو حرفی و ... از مدرک حذف می‌شوند؛

۳ – ریشه یابی واژه‌ها؛ این مرحله با تبدیل شکل‌های گوناگون یک اصطلاح به ریشه آن، موجب کاهش تعداد اصطلاحات می‌شود؛

۴ – حذف واژه‌هایی با بسامد بسیار کم؛

۵ – انتخاب پر بسامدترین واژه‌ها به عنوان واژه‌های کلیدی و مهم؛

۶ – شمارش تعداد دفعاتی که این واژه‌های مهم در مدرک ظاهر شده اند، به منظور محاسبه TF؛

۷ – تعیین تعداد کل مدارک موجود در پایگاه؛

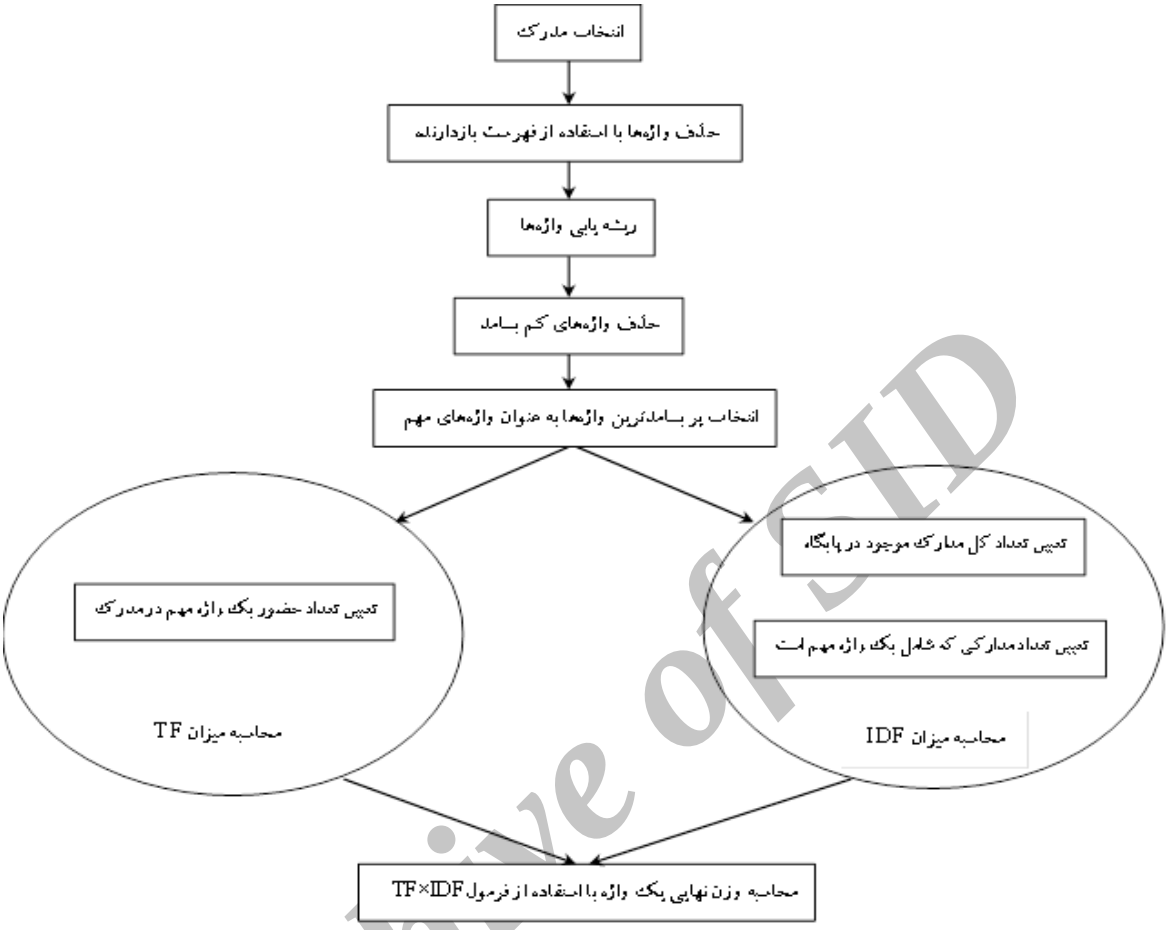
۸ – شمارش تعداد مدارکی که هر واژه مهم درآن ظاهر شده است؛

۹ – محاسبه میزان IDF با استفاده از اعداد به دست آمده از مرحله ۷ و ۸ ...؛

۱۰ – استفاده از اعداد به دست آمده از مرحله ۶ و ۹ به منظور محاسبه وزن هر واژه مهم به صورت مجزا و با استفاده از فرمول روش ترکیبی TF×IDF.

1. context- based term weighting
2. Stop List

مراحل وزن‌دهی به یک واژه در شکل ذیل نشان داده شده اند:



شکل ۱. فرایند وزن‌دهی خودکار به اصطلاحات

نقش وزن‌دهی در جامعیت و مانعیت

عملکرد اصلی یک نظام وزن‌دهی افزایش میزان بازیابی اطلاعات است. بازیابی موثر به دو عامل اصلی بستگی دارد (Ropero, et.al, 2012; Salton, Buckley, 1988):

۱. مرتبط ترین آیتم‌ها به نیاز کاربر باید بازیابی شوند.

۲. آیتم‌های غیر از آن باید کنار زده شوند؛ به عبارتی آیتم‌های نامرتبط در یک مجموعه نباید بازیابی شوند.

عامل اول را جامعیت[1] و عامل دوم را مانعیت[2] می‌نامند.

جامعیت هنگامی به حد مطلوب می‌رسد که از اصطلاح‌هایی با بسامد بالا که در بسیاری از مدارک مجموعه ظاهر می‌شوند استفاده شود. انتظار می‌رود که استفاده از این اصطلاح موجب بازیابی تعداد زیادی از مدارک که شامل مدارک مرتبط بسیار می‌باشد، شود. اما مانعیت هنگامی به حد مطلوب می‌رسد که اصطلاح‌هایی خاص که قادر هستند تعداد کمی از آیتم‌های مرتبط را از توده ی آیتم‌های نامرتبط مجزا کنند استفاده شوند. بر این اساس روش وزن‌دهی TF که بر اساس بسامد اصطلاح می‌باشد موجب

1. Recall
2. Precision

بهبود جامعیت می‌شود و روش وزن‌دهی IDF که بر اساس بسامد معکوس مدرک در کل مجموعه است موجب بهبود مانعیت می‌شود. از آنجا که در عمل همواره سطح منطقی از جامعیت بدون ایجاد مانعیت خیلی پایین، مورد توجه قرار می‌گیرد؛ پس باید بینابین این دو شیوه مورد استفاده قرار گیرد که هم مانعیت و هم جامعیت را در برگیرد. از این رو، برای بهبود جامعیت و مانعیت باید از روش ترکیبی TF×IDF استفاده شود که نشان می‌دهد بهترین اصطلاح‌ها آنهایی هستند که در مدرک بسامد بالا داشته اند اما دارای بسامد پایین در کل مجموعه هستند (Ropero, et.al, 2012; Salton, Buckley, 1988).

جمع‌بندی

بازیابی اطلاعات بطور گسترده در زندگی روزمره مورد استفاده قرار می‌گیرد. آنچه برای طراحان سیستم‌های بازیابی اطلاعات از اهمیت بالایی برخوردار است، افزایش کارایی و بهبود عملکرد این سیستم‌ها در بازیابی مدارک مرتبط با پرسش کاربران می‌باشد. همانگونه که اشاره شد، یکی از راه‌های افزایش بهره‌وری سیستم‌های بازیابی اطلاعات، استفاده از طرح‌های وزن‌دهی به اصطلاح‌ها می‌باشد. در این طرح‌ها به کلیدواژه‌ها وزن‌هایی بر اساس اهمیت اختصاص می یابد که میزان سودمندی آنها در شناساندن موضوع یک مدرک و میزان ربط آن با پرسش کاربر را نشان می‌دهد. به عبارتی واژگانی که در مدرک تأثیر بیشتری دارند شناسایی می‌شوند. بنابراین وزن‌دهی موجب می‌شود کاربر بتواند از میزان ارتباط مدارک بازیابی شده با پرسش خود آگاهی یابد، زیرا بر اثر وزن‌دهی، مدارک بازیابی شده بر اساس درجه مشابهت بین عبارت پرسش و مدارک، مورد استفاده قرار می‌گیرند. همچنین منجر به رتبه‌بندی مدارک بازیابی شده می‌شود و در نتیجه کاربر می‌تواند در مورد نتایج ارائه‌شده قضاوت کند. در مجموع وزن‌دهی اصطلاحات عملکرد بازیابی اطلاعات را ارتقا می‌بخشد. در سال‌های اخیر چگونگی تخصیص وزن مناسب به اصطلاح‌ها مورد توجه قرار گرفته است. پژوهش‌ها نشان داده‌اند که طرح‌های پیچیده وزن‌دهی نسبت به شیوه‌های ساده که صرفاً بر اساس حضور هر اصطلاح در متن می‌باشد، از عملکرد بهتری در بازیابی اطلاعات برخوردارند. بنابراین پژوهشگران در وزن‌دهی به اصطلاح‌ها، در جستجوی شیوه‌هایی با کارایی بالاتر می باشند که در عمل موجب بهبود بازیابی مدارک می‌شوند.

همواره شاهد ابداع روش‌های پیچیده ای از وزن‌دهی هستیم که توجه پژوهشگران را به خود جلب کرده است. اما با وجود تمرکز زیاد بر تحقیقات در زمینه روش‌های وزن‌دهی، شایان ذکر است که وزن‌دهی به اصطلاح‌ها تنها راه افزایش و بهبود کارکرد سیستم‌های بازیابی اطلاعات نیست و مجموعه ای از عوامل در بهبود این سیستم‌ها تأثیر گذارند.

منابع

پولیت، ا. استون (۱۳۸۳). نظام‌های ذخیره و بازیابی اطلاعات: خاستگاه، توسعه و کاربردها. ترجمه محمد حسین دیانی، جعفر مهراد. شیراز: کتابخانه منطقه ای علوم و تکنولوژی شیراز.

حسن زاده، محمد (۱۳۸۳). «تأثیر مدل‌های بازیابی اطلاعات بر میزان ربط». *اطلاع شناسی*. ۲ (۱)، ۶۴-۸۹.

کوکبی، مرتضی (۱۳۸۷). «استفاده از ضریب وزنی در سرعنوان‌های موضوعی در راستای بهبود بازیابی اطلاعات». فصلنامه کتابداری و اطلاع‌رسانی. *۱۱* (۴۱)، ۲۴۳-۲۵۸.

گراسمن، دیوید ا. ؛ فریدر، افیر (۱۳۸۴). بازیابی اطلاعات: الگوریتم‌*ها و روش‌های اکتشافی*. ترجمه جعفر مهراد و سارا کلینی. مشهد: کتابخانه رایانه‌ای ؛ شیراز: کتابخانه منطقه‌ای علوم و تکنولوژی.

لارج، آندرو؛ تد، لوسی؛ هارتلی، ریچارد (۱۳۸۲). جستجوی اطلاعات در عصر اطلاعات. ترجمه زاهد بیگدلی، ویراسته زهیر حیاتی. تهران: کتابدار.

لنکستر، اف. دبلیو (۱۳۸۲). نمایه سازی و چکیده نویسی: مبانی نظری و عملی. ترجمه عباس گیلوری. تهران: چاپار.

مهراد، جعفر؛ کلینی، سارا (۱۳۸۶) . « بررسی مدل فضا برداری در بازیابی اطلاعات». *فصلنامه کتابداری و اطلاع‌رسانی. ۱۰* (۳۸)، ۱۹۷-۲۱۰.

میدو، چارلز تی، بویس، برت آر، کرفت، دونالد اچ، باری، کارول (۱۳۹۰). نظام‌های بازیابی اطلاعات متنی. ترجمه نجلا حریری. تهران: چاپار.

Bisht, Raj Kishor, Srivastava, Garima, Dhami, H. S. (2010). "Term weighting using term dependence". *International Journal of Computer Applications, 3*(11), 1-3.

Cheng, Juan (2006). *"A single document-based term weighting scheme by supporting terms*". (Master's Thesis), Utah State University.

Drucker, Harris, Shahrary, Behzad, Gibbon, David C. (2002). "Support vector machines: relevance feedback and information retrieval*"*. *Information Processing and Management*, *38*(3), 305-323.

Ernandes,Marco et al. (2007). *"An adaptive context-based algorithm for TermWeighting : application to Single-Word Question Answering*". Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007) , AAAI Press, Jan. 2007.

Hong, Min, et al. (2005). *"Integrated term weighting, visualization, and user interface development for Bioinformation retrieval*". *Springer-Verlag* : 673-682.

Jones, Karen Sparck (2004). "IDF term weighting and IR research lessons". *Journal of Documentation, 60* (5), 521-523.

Kang, Bo-Yeong, Kim, Dae-Won, Lee, Sang-Jo (2005). "Exploiting concept clusters for content-based information retrieval*"*. *Information Sciences, 170*(2-4), 443–462.

Kowalski, G. (1998). *Information Retrieval Systems: Theory and Implementation*. Kluwer, Boston.

Nanas, Nikolaos, Uren, Victoria, Roeck, Anne De. (2003). *"A Comparative Study of Term Weighting Methods for Information Filtering*". Proceedings of International Workshop on Database and Expert Systems Applications DEXA.04, 13-17.

Radecki, Tadeusz (1983). "Generalized Boolean methods of information retrieval". *Int. J. Man-Machine Studies, 18*(5), 407-439.

Raju, N. V. Ganapathi, Sukavasi, Bhavya, Chava, Sai Rama Krishna (2011). "An application of statistical indexing for searching and ranking of documents- A case study on Telugu script". *International Journal of Computer Applications*, *28*(3), 22-27.

Robertson, Stephen (2004). "Understanding inverse document frequency: On theoretical arguments for IDF". *Journal of Documentation, 60*(5), 503-520.

Ropero, Jorge (2012). "A Fuzzy Logic intelligent agent for Information Extraction: Introducing a new Fuzzy Logic-based term weighting scheme". *Expert Systems with Applications, 39*(4), 4567–4581.

Ruch, P., Baud, R., Geissbuhler, A. (2002). "Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records". *International Journal of Medical Informatics*, *67*(1-3), 75-83.

Salton, Gerard, Buckley, Christopher (1988). "Term-weighting approaches in automatic text retrieval". *Information Processing and management. 24*(5), 513-52.

Trotman, Andrew (2005). "Choosing document structure weights". *Information Processing and Management. 41*(2), 243–264.

Witschel, Hans Friedrich (2008). "Global term weights in distributed environments" *Information Processing and Management, 44*(3), 1049–1061.