ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال پنجم)
شماره ۱۷ / بهار ۱۳۹۹ / ص ۱۰۸– ۹۹

بررسی و مقایسه ویژگی‌های مستخرج از کلیدواژه‌های مقالات فارسی

نیلوفر مظفری
استادیار مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

نام نویسنده مسئول:
نیلوفر مظفری

تاریخ دریافت: ۱۳۹۹/۱/۲۳
تاریخ پذیرش: ۱۳۹۹/۳/۲۱

چکیده

کلیدواژه‌ها به کلماتی اطلاق می‌گردند که از متنی طولانی استخراج شده و گویای محتوای آن متن هستند. در واقع خواننده با مطالعه آنها به محتوای اصلی متن و مفاهیم و موضوعات مورد توجه نویسنده، پی می‌برد. علاوه بر این، کلیدواژه‌ها در فرآیند نمایه کردن یک متن علمی نقش بسزایی دارند. بنابراین انتخاب درست کلیدواژه‌ها به نمایه سازی استاندارد متون در فضای مجازی کمک می‌کند و دسترسی مخاطبان را به اثر پژوهشی تسهیل می‌کند. کلیدواژه ها باید با واژه‌های اصلی عنوان و مساله تحقیق و در حدامکان با سرفصل‌ها تناسب داشته باشند و میانگین تعداد آنها بین ۵ تا ۷ کلمه است. مهمترین کلیدواژه‌های هر مقاله، مربوط به متغیرهایی از متن هستند که طی پژوهش، توسط محققین اندازه‌گیری شده‌اند.

نظر به اهمیت کلیدواژه‌ها در مقالات علمی، در این پروژه به تحلیل و بررسی آنها پرداخته و ویژگی‌هایی مبتنی بر گروه اصلی نشریات وزارت علوم از این کلمات استخراج می‌گردد. سپس این ویژگی‌ها تحلیل و با یکدیگر مقایسه می‌شوند. لازم به ذکر است که از این ویژگی‌ها می-توان در ساخت یک هستان شناسی بهره برد و همچنین به سامانه‌های بازیابی اطلاعات در جهت رسیدن به افزایش کاربرپسندتر بودن کمک نمود..

واژگان کلیدی: کلیدواژه، ویژگی، مقالات علمی، برچسب‌دهی موضوعی.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

در گذشته اطلاعات روی رسانه‌های فیزیکی مانند کتاب ذخیره می‌شد. امروزه با گسترش و توسعه اطلاعات الکترونیکی از یک طرف و رشد سریع اطلاعات الکترونیکی از طرف دیگر، با حجم اطلاعات قابل دسترس بسیاری روبرو شده‌ایم. در واقع می-توان گفت که رشد سریع دانش و انتشار انبوه اسناد کتابی و یا غیرکتابی بویژه پس از جنگ جهانی دوم، بزرگترین واقعه تاریخ علوم کتابداری و اطلاع‌رسانی است. این امر منجر به ایجاد روش‌های جدیدی برای تجزیه و تحلیل اسناد شد، زیرا روش‌های قبلی دیگر کافی نبودند [۱].

وظیفه مراکز اسناد یا کتابخانه‌ها این است که مدارک و اسناد[1] را طوری سازماندهی و ذخیره نمایند تا کاربران بتوانند به سهولت اطلاعات لازم را از میان آنها بازیابی نمایند. لازم به ذکر است که مدرک یا سند هر چیز چاپی یا غیرچاپی است که قابلیت فهرست و یا نمایه شدن را داشته باشد [۲].

همانطور که ورود رایانه و استفاده از نرم‌افزارهای کتابخانه‌ای در حوزه کتابداری و اطلاع‌رسانی موجب افزایش سرعت و دقت در دسترسی به اطلاعات باشد، بهره‌گیری از نظام‌های هوشمند نیز می‌تواند در بسیاری از فعالیت‌های کتابخانه‌ها موثر باشد.

یکی از چالش‌های پیش‌رو در جهت پردازش اسناد، استخراج ویژگی‌ها از کلیدواژه‌های موجود در اسناد می‌باشد که کابردهای بسیار زیادی در حوزه‌های مختلف مانند ایجاد هستان‌شناسی[2]ها [۳]، خوشه‌بندی[3] [۴] و طبقه‌بندی[4] [۵] اسناد دارد.

کلیدواژه‌ها که در نمایه‌سازی مقالات نیز بسیار مورد توجه قرار دارند، به کلماتی گفته می‌شوند که از متنی طولانی گرفته شده و گویای محتوای آن متن هستند. در واقع خواننده با مطالعه آنها به محتوای اصلی متن و مفاهیم و موضوعات مورد توجه نویسنده، پی می‌برد. معمولاً دو روش اصلی برای انتخاب کلیدواژه ها وجود دارد. روش اول که به انتخاب کلیدواژه‌های کنترل شده معروف است، به کمک اصطلاح نامه ها صورت می‌گیرد. در روش دوم که تحت عنوان انتخاب آزاد شناخته می‌شود، براساس تصمیم نویسنده صورت می‌گیرد. در هریک از این دو روش لازم است ترکیبی از واژگان اعم و اخص که بیشترین و نزدیک ترین رابطهٔ معنایی را با محتوای مقاله دارد، انتخاب شود. انتخاب کلیدواژهٔ مناسب در مقاله احتمال بازیابی بیشتر مقاله را توسط دیگران در آینده افزایش می‌دهد و درنتیجه به دریافت استناد بیشتر به مقاله کمک می‌کند. درمجموع، کلیدواژه ها باید با موضوع اصلی مقاله تناسب کافی داشته باشند.

هستان‌شناسی‌ها به عنوان یکی از مهمترین فناوری‌های وب معنایی، از جمله دستاوردهای هوش مصنوعی هستند که علاوه بر داشتن نقش کلیدی در تحقق چشم‌انداز وب معنایی، کاربردهای مختلفی نیز در بهبود کیفیت بازیابی اطلاعات مبتنی بر کلیدواژه داشتند [۶]. در واقع آنها با تعریف مفاهیم اصلی یک حوزه موضوعی، مبادرت به تعریف یک واژگان مشترک می‌کنند که به واسطه آن تعامل میان کامپیوتر و انسان میسر گردد. سپس با تعریف روابطی میان این واژه‌ها، امکان استنتاج معنایی و غنی‌سازی رسایی معنایی را برای کاربردهای مختلف از جمله نمایه‌سازی و پرسش‌های جستجو فراهم می‌کنند [۷].

خوشه‌بندی به عنوان یکی از مهمترین تکنیک‌ها در حوزه یادگیری بدون ناظر[5] ، عملیات تبدیل داده‌ها به گروه‌ها را انجام می‌دهد. این گروه‌بندی طوری انجام می‌شود که داده‌هایی که در یک گروه قرار می‌گیرند بیشترین شباهت را با یکدیگر داشته و با داده‌های سایر گروه‌ها کمترین میزان شباهت را دارند. در این صورت به هر کدام از این گروه‌ها یک خوشه[6] و به این عملیات، خوشه‌بندی گفته می‌شود. در خوشه‌بندی داده‌های متنی، ابتدا از هر سند، ویژگی‌هایی استخراج شده که این ویژگی‌ها مبتنی بر کلیدواژه هستند و کارایی نهایی خوشه‌بندی به شدت به انتخاب این ویژگی‌های اولیه وابسته است [۸].

انتساب اسناد متنی به موضوعات از پیش تعیین شده به منظور طبقه‌بندی خودکار متون در ده سال اخیر توجهات را به خود جلب کرده است. روش اصلی در طبقه‌بندی خودکار موضوعی متون، یادگیری ماشینی[7] است و از متداولترین روشهای مورد

---

[1] document
[2] ontology
[3] clustering
[4] classification
[5] Unsupervised learning
[6] cluster
[7] Machine learning

ISSN: 2745-3673
http://www.Rcsj.ir

استفاده در دسته‌بندی متون می‌توان به روش‌های درخت‌های تصمیم‌گیری[8]، نزدیکترین همسایه، ماشین‌های بردار پشتیبان[9]، شبکه‌های عصبی[10]، منطق فازی[11] و بیزین ساده اشاره کرد [۵]، [۹]، [۱۰] و [۱۱].

در حوزه بازیابی اطلاعات وتحلیل داده‌های آماری، برچسب زنی موضوعی متون امری مهم و کاربردی به شمار میرود. متخصصان حوزه اطلاع‌رسانی برای سازماندهی موضوعی مدارک به منظور بازنمون موضوعیت آنها و تسهیل ارتباط میان زبان کاربر و نظام بازیابی اطلاعات، از برچسب‌های موضوعی استفاده می‌کنند. این بسته‌های اطلاعاتی کوچک که پیشتر به صورت دستی (با دخالت انسان) و یا خودکار (مانند نمایه‌سازی استخراجی) مورد ارزیابی و داوری قرار گرفته‌اند تا حامل بیشترین بار محتوایی مرتبط با مدرک باشند، به منزله میانبرهای شناختی برای بازنمون محتوای موضوعی مدارک، مخاطب را بدون نیاز به اطلاعات زیاد و تفکر موشکافانه، به سرعت به ارزیابی و نتیجه گیری می‌رسانند.

جستجوی موضوعی در پایگاه های اطلاعاتی به واسطهٔ برچسب‌های موضوعی اختصاص یافته، اعم از کلیدواژه‌ها و یا توصیفگرها، صورت می پذیرد. با این تفاسیر، کلیدواژه‌ها یا اصطلاح‌هایی که به منزله شناساگر و یا توصیفگر مدارک درکلیه خدمات نمایه‌سازی و چکیده‌نویسی کاربرد دارد، همراه با سرعنوان‌های موضوعی و یا فراداده‌های موضوعی اختصاص یافته به مدارک تحت وب، همگی شرایط برچسب بودن را دارا هستند. از نظر کاربر نهایی، مدارک بازیابی شده با حوزه موضوعی مورد جستجو مرتبط خواهند بود. کاربر میانی نیز با مطالعه متن تشخیص می دهد که مدرک در کدام حوزه موضوعی جای می‌گیرد. در صورتیکه با وجود مجموعه برچسب‌های موضوعی، بدون آنکه به مطالعه عمیق‌تر چکیده و یا حتی متن اصلی اثر نیاز باشد، این امکان فراهم می‌گردد تا دیدی کلی نسبت به اثر بدست آورد. البته، نباید فراموش کرد که برچسب‌های موضوعی، فاقد سوگیری ارزشی مثبت یا منفی برای مدارک بوده و در نهایت ارزشگذاری مدارک توسط هر کاربر و با توجه به ربط شخصی وی، متفاوت خواهد بود و از آنجا که این درک یک ارزیابی ذهنی است، برای افراد مختلف می‌تواند متفاوت باشد [۱۲].

نظر به اهمیت کلیدواژه‌ها در حوزه‌های مختلف، این مقاله به بررسی ویژگی‌های مختلفی که می‌توان از آنها استخراج کرد، پرداخته است. علاوه بر آن، با استفاده از برچسب‌دهی که به کلیدواژه‌ها بر اساس هر کدام از ویژگی‌ها می‌دهد، می‌توان حوزه موضوعی یک متن را نیز استخراج نمود. در ادامه مروری بر روش‌هایی که به نحوی به برچسب‌زنی داده‌ها می‌پردازند، ارائه می-شود. سپس روش پیشنهادی در انتخاب ویژگی مبتنی بر کلیدواژه بر اساس سرعنوان‌های موضوعی وزارت علوم و در نهایت شبیه‌سازی و نتیجه‌گیری بیان می‌گردد.

کارهای پیشین

نظریه برچسب‌زنی برای اولین بار توسط «هاوارد بکر» در سال ۱۹۶۳ مطرح گردید. این نظریه در واقع یکی از نظریه های حوزه علوم اجتماعی و روانشناسی و برای کاربرد در زمینه‌های بزهکاری و جرم شناسی مطرح شده است. در حوزه سازماندهی اطلاعات و برچسب‌گذاری موضوعی نیز از برچسبها به منزلهٔ روشی برای تبادل اطلاعات کاربران و نظام اطلاعاتی در درجه اول و ایجاد و جهت‌دهی به نگرش در باره موضوعیت مدارک بازیابی شده (به صورت مستقیم یا غیر مستقیم) در درجه دوم استفاده می‌شود. برچسبها دراین حوزه، رسمی، غنی از اطلاعات و در قالب بیان های لغوی اما فاقد سوگیری ارزشی می باشند.

لیو و یانگ دسته بندی متون را با استفاده از بردارهای فراوانی ریشه کلمات انجام دادند. این دسته بندی در سال ۱۹۹۸ با استفاده از ماشین بردار پشتیبان توسط جاچیمز مورد مطالعه قرار گرفت [۱۳]. بلگاردا روش آنالیز معنایی پنهان (LSA) را برای دسته بندی به کار برد [۱۴]. وود و گوداون از شبکه های عصبی هیبرید به منظور دسته بندی متون استفاده کردند [۱۵]. ترکولا آنالیز تمایزی خطی را در دسته بندی به کارگرفت [۱۶]. بلی و همکاران روش تخصیص دیریکلهٔ پنهان (LDA) را برای مدلسازی متون پیشنهاد دادند و از آن در دسته بندی متون نیز استفاده کردند [۱۷]. گواندانگ و همکاران روش تحلیل معنایی پنهان احتمالاتی (PLSA) را برای دسته بندی صفحات وب به کارگرفتند [۱۸]. عملکرد طبقه بندی متن از طریق یک اصلاح

[8] Decision tree
[9] SVM
[10] Neural Network
[11] Fuzzy logic

ISSN: 2745-3673
http://www.Rcsj.ir

نامه مبتنی بر پیکره وردنت، با به کارگیری الگوریتم نزدیکترین همسایگی و شبکه عصبی پس انتشار الگوریتم را بهبود دادند [۱۹].

عرب سرخی و فیلی یک روش دسته بندی با استفاده از بردارهای فراوانی ریشهٔ کلمات و الگوریتم بیزین ساده پیشنهاد داده‌اند و با ترکیب روش بیزین با ایده نگهداری کلمات همنشین، روش خود را بهبود بخشیدند [۲۰]. حاجی حسینی و الماس گنج روش نظارت برای دسته بندی متون فارسی با استفاده از تحلیل معنایی پنهان (LSA) را پیشنهاد دادند [۲۱]. این روش بردارهایی را در یک فضای برداری کاهش یافته برای هر متن در اختیار قرار می دهد که با استفاده از آنها شبکهٔ عصبی برای آموزش و تعیین دستهٔ مربوط به متون جدید تشکیل می گردد. درپژوهشی دیگر، با استفاده از روش بهره جویی از گنج واژه و انتخاب ویژگی دو مرحله ای به دسته بندی متون فارسی پرداخته اند. قویدل و همکاران نیز برای دسته بندی متون از روش فاصله یابی در فضای بردار بسامدی بهره گرفته اند [۲۲].

پژوهش‌های دیگری هم در سال‌های اخیر برای مدل‌سازی موضوعی ارائه شده است که ارتباط میان واژگان را در سطحی محلی‌تر بررسی می‌کنند. نخستین روشی که در این دسته ارائه گردید و به نحوی الهام‌بخش دیگر روش‌های این دسته بود، BTM می‌باشد. در این روش فرض می‌شود که هر واژه علاوه بر موضوع خود، وابسته به موضوع واژه پیشین خود نیز هست [۲۳]. گریفیس و همکارانش فرض کردند که هر واژه توسط یک موضوع و یا توسط واژه پیشینش تولید می‌گردد که برای انتخاب یکی از این دو حالت، از یک متغیر برنولی استفاده کردند [۲۴]. تعمیمی بر این روش، توسط ونگ و همکارانش انجام گردید که در آن هر واژه بر مبنای موضوع خود می‌تواند تصمیم بگیرد که آیا با واژه قبلی یک ترکیب را تشکیل دهد یا خیر [۲۵]. در پژوهشی دیگر فرض مشابهی در نظر گرفته شد و علاوه بر آن، فرض شد که یک سلسله مراتب از موضوعات وجود دارد و هر واژه، مسیری مشخص را در این سلسله مراتب طی می‌کند تا توسط یک موضوع خاص تولید گردد [۲۶]. یک مدل موضوعی باناظر در ترکیب با مدل زبانی بایگرام توسط جمیل و همکاران ارائه شد [۲۷]. دلیل استفاده از از بایگرام در این مدل-ها، تنک بودن داده می‌باشد.

عمده این روش‌ها مبتنی بر ویژگی‌هایی هستند که به نحوی از کلیدواژه‌ها استخراج می‌شود. بنابراین در این پژوهش به بررسی ویژگی‌های مختلفی که از کلیدواژه‌ها استخراج می‌شود می‌پردازیم و ارتباط آنها را با سرعنوان‌های موضوعی نشریات وزارت علوم بررسی می‌نماییم.

روش پیشنهادی

روش پیشنهادی چهار ویژگی را برای هر کلیدواژه پیشنهاد می‌دهد که همگی بر اساس گروه نشریات وزارت علوم[12] می-باشند. جدول ۱ لیست این گروه‌ها را نشان می‌دهد. در این پژوهش، به ازای هر گروه موجود در نشریه وزارت علوم ۱۰ نشریه مورد بررسی قرار گرفت و به ازای هر نشریه ۵۰ مقاله موجود در آن پردازش گردید.

جدول ۱: گروه‌های موجود در نشریات وزارت علوم

ردیف	گروه
۱	علوم انسانی
۲	فنی و مهندسی
۳	کشاورزی و منابع طبیعی
۴	دامپزشکی
۵	علوم پایه
۶	هنر و معماری

[12] https://journals.msrt.ir/

ISSN: 2745-3673
http://www.Rcsj.ir

ویژگی اول همان frequency یا تکرار هر کلیدواژه در هر گروه است. برای انجام این کار، به ازای هر گروه نشریات موجود در آن و به ازای هر نشریه، مقالات و به ازای هر مقاله، کلیدواژه‌های آن مقاله بررسی شده و بر اساس تعداد تکرار هر کلیدواژه در هر گروه یک وزن به آن کلیدواژه انتساب داده می‌شود. این ویژگی را $fr_{i,j}$ می‌نامیم که i و j به ترتیب نشان‌دهنده کلیدواژه و گروه می‌باشند که این گروه‌ها بر اساس جدول ۱ مشخص می‌گردند. به عنوان مثال $fr_{\text{نظامی ,انسانی علوم}} = 5$ به معنی تکرار کلیدواژه "نظامی" در گروه "علوم انسانی" به اندازه ۵ بار می‌باشد.

مشکلی که این ویژگی دارد، این است که برای هر کلیدواژه، وزن بالایی به گروه‌هایی می‌دهد که حاوی کلیدواژه‌های زیادی باشند. برای حل این مشکل، ویژگی دوم و یا همان فرکانس نرمال شده که با NF نمایش می‌دهیم، استفاده می‌نماییم که فرمول (۱) موید این امر است.

$$NF_{i,j} = \frac{fr_{i,j}}{\max fr_{i,j}} \qquad (۱)$$

در این فرمول، $fr_{i,j}$ نشاندهنده فرکانس تکرار کلیدواژه i در گروه j است و مخرج کسر، ماکزیمم تکرار کلیدواژه در گروه j است.

ویژگی سوم که با LN نمایش می‌دهیم، به ازای هر کلیدواژه مشخص می‌کنند که چقدر این کلیدواژه در گروه‌های مختلفی وجود دارد. طبیعتا هر چقدر که یک کلیدواژه در تعداد گروه‌های کمتری وجود داشته باشد، به معنی متمایز بودن آن کلیدواژه برای گروه مورد نظر است. فرمول (۲) نحوه استخراج این ویژگی را از هر کلیدواژه نشان می‌دهد.

$$LN_i = \log_e(s/s_i) \qquad (۲)$$

در این فرمول، S نشاندهنده تعداد کل گروه‌ها و s_i تعداد گروه‌های اصلی را نشان می‌دهد که کلیدواژه iم در آن وجود دارد و e عدد نپر است.

ویژگی چهارم، ترکیب ویژگی دوم و سوم می‌باشد. بنابراین از مزایای هر دو ارث می‌برد. فرمول (۳) نحوه استخراج این ویژگی را از هر کلیدواژه مشخص می‌کند.

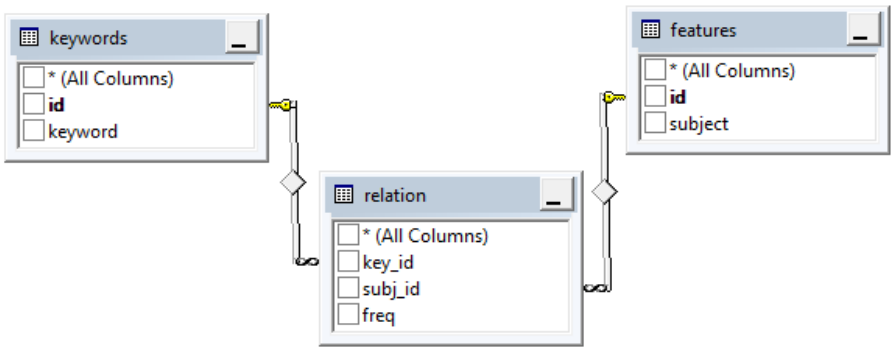
$$F_{i,j} = \mathrm{NF}_{i,j} \times \mathrm{LN}_i \qquad (۳)$$

در این فرمول، $F_{i,j}$ نشان‌دهنده ویژگی استخراج شده برای کلیدواژه i در گروه i است و $\mathrm{NF}_{i,j}$ ویژگی دوم و LN_i ویژگی سوم معرفی شده در این مقاله را نشان می‌دهد.

شبیه‌سازی

برای شبیه‌سازی ویژگی‌های تعریف شده در بخش قبل ، ابتدا تمامی کلیدواژه‌ها پردازش شده و این کلمات به همراه تعداد تکرار در گروه‌های مختلف در یک پایگاه داده ذخیره شدند. شکل ۱ شمایی از جداول استفاده شده در پایگاه داده این پژوهش را نشان می‌دهد.

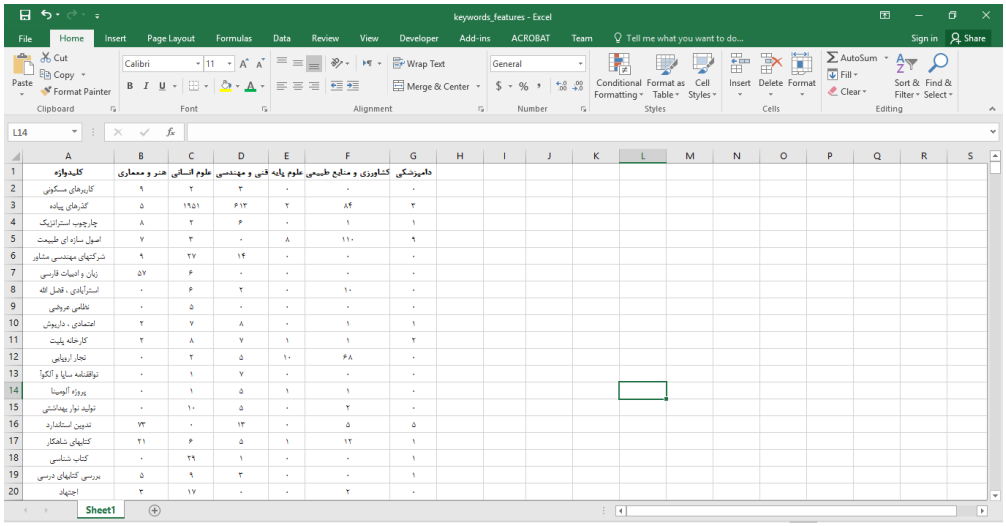
ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۱

به ازای هر کلیدواژه، در صورتی که این کلیدواژه قبلا در جدول kewords وجود نداشته باشد، یک سطر به این جدول اضافه می‌گردد و بعد از مشخص شدن گروه نشریه‌ای که مقاله حاوی این کلیدواژه در آن قرار دارد، اطلاعات جدول relation به‌روز می‌گردد.

در مرحله بعدی این جدول به Excel، export گردید. شکل ۲ کلیدواژه‌ها را به همراه گروه‌ها نشان می‌دهد. این جدول شامل ۷ ستون می‌باشد. ستون اول، کلیدواژه مربوطه را نشان می‌دهد و ستون‌های دوم تا هشتم، ۶ موضوع اصلی و یا همان گروه‌های اصلی نشریات وزارت علوم به نام‌های "هنر و معماری"، "علوم انسانی"، "فنی و مهندسی"، "علوم پایه"، "کشاورزی و منابع طبیعی" و "دامپزشکی"را مشخص می‌نماید. سطرهای این جدول، کلیدواژه‌ها را نشان می‌دهد و ستون‌های مربوطه نشان-دهنده تعداد تکرار کلیدواژه در آن موضوع اصلی و یا گروه‌های اصلی نشریات وزارت علوم است. به عنوان مثال کلیدواژه "کاربرهای مسکونی" در ۹ نشریه با موضوع اصلی "هنر و معماری"، در ۲ نشریه با موضوع اصلی "علوم انسانی" و ۳ نشریه با موضوع اصلی "فنی و مهندسی" تکرار شده است.



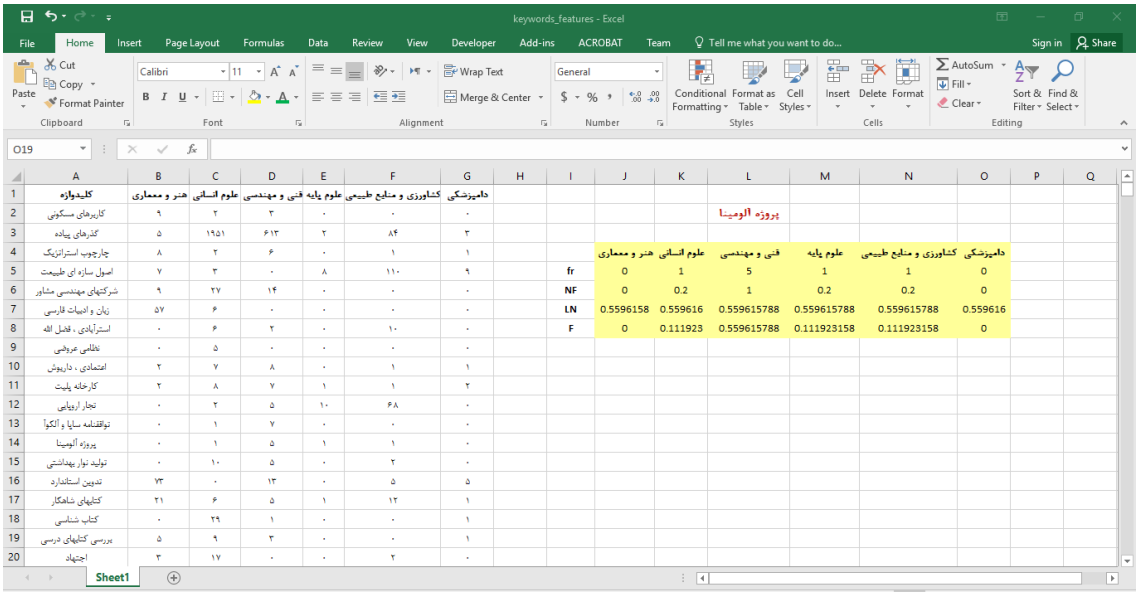
شکل ۲: کلیدواژه‌ها به همراه موضوعات اصلی بعد از export به فایل Excel

شکل ۳ چهار ویژگی ارائه شده در این مقاله را نشان می‌دهد. لازم به ذکر است که این پژوهش با زبان VBA[۱۳] نوشته شده است و به همراه فایل اکسل قابل دسترسی است. (شکل ۴ نمونه‌ای از کد نوشته شده در این زبان برای این پژوهش را نشان می‌دهد)

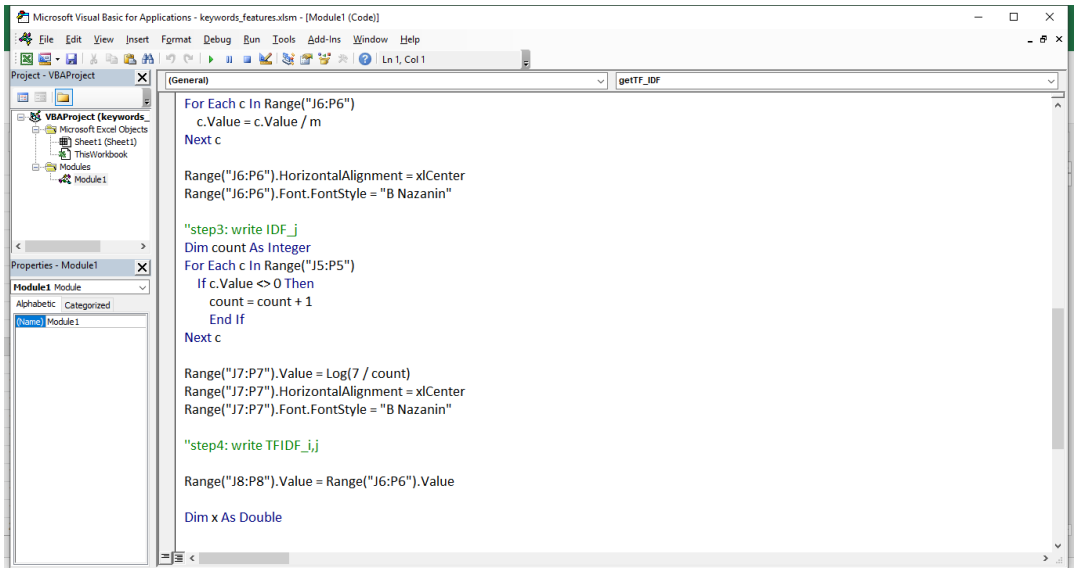
[13] Visual Basic for Applications

ISSN: 2745-3673
http://www.Rcsj.ir

کاربر با انتخاب هر کلیدواژه و فشردن دکمه‌های میانبر ctrl+shift+Q می‌تواند ویژگی‌های آن را مشاهده نماید. به عنوان نمونه، چهار ویژگی استخراج شده برای کلیدواژه "پروژه آلومینا" در شکل ۳ نشان داده شده است.



شکل ۳: ویژگی استخراج شده از کلیدواژه‌ها



شکل ۴: قسمتی از کد مربوط به استخراج ویژگی با VBA

هر کدام از چهار ویژگی ارائه شده در این مقاله بسته به کاربرد آن می‌توانند مورد استفاده قرار بگیرند. به عنوان مثال از این ویژگی‌ها می‌توان در فرآیند ساخت یک هستان‌شناسی بهره برد و آنها را به عنوان predicate به object نسبت داد. علاوه بر این، می‌توان از این ویژگی‌ها برای مشخص کردن گروه یک مقاله نیز استفاده کرد. در ادامه مقایسه‌ای از این ویژگی‌ها آورده می‌شود.

ویژگی fr یک ویژگی سریع و ساده برای کلیدواژه‌ها است. در کاربردهایی که نیاز به استخراج ویژگی برای کلیدواژه‌ها در یک زمان کوتاه است، این ویژگی می‌تواند موثر واقع گردد. علاوه بر این، در بعضی از کاربردها، تعداد تکرار دقیق هر کلیدواژه در

ISSN: 2745-3673
http://www.Rcsj.ir

هر گروه مورد نیاز می‌باشد. اشکالی که این ویژگی دارد این است که به گروه‌هایی که تعداد نشریه زیادی در آن قرار دارد و یا نشریاتی که مقالات زیادی دارند و یا مقالاتی که تعداد کلیدواژه زیادی در آن وجود دارد، وزن بیشتری می‌دهد.

ویژگی NF این مشکل را برطرف می‌کند چون بر ماکزیمم تعداد تکرار کلیدواژه در آن گروه تقسیم می‌کند که این امر موجب نرمال‌سازی می‌شود.

در کاربردهایی که نیاز به خوشه‌بندی و یا طبقه‌بندی مقالات با توجه به کلیدواژه‌ها می‌باشد، ویژگی NF با مشکل روبرو می‌شود؛ چرا که در این کاربردها نیاز است که طوری به کلیدواژه‌های مقالات، ویژگی انتساب دهیم که در نهایت دقت خوشه-بندی و یا طبقه‌بندی بیشینه گردد. این امر نیاز به این است که بتوان با استفاده از کلیدواژه تصمیم درستی در مورد کلاس یا خوشه مقاله کرد. برای توضیح بیشتر کلیدواژه نظامی عروضی در شکل ۲ فقط در گروه علوم انسانی وجود دارد. بنابراین، این کلیدواژه یک کلیدواژه بسیار خوب برای گروه علوم انسانی می‌باشد. در صورتی که یک مقاله حاوی این کلیدواژه باشد، با احتمال بسیار زیاد این مقاله مربوط به گروه علوم انسانی است. در این شکل، کلیدواژه "کارخانه پلیت" در تمامی گروه‌های هنر و معماری، علوم انسانی، فنی و مهندسی، علوم پایه، کشاورزی و منابع طبیعی، دامپزشکی وجود دارد. پس این کلیدواژه نمی‌تواند معرف خوبی برای هیچکدام از گروه‌ها باشد و با دیدن این کلیدواژه در یک مقاله، نمی‌توان در مورد گروه آن اظهار نظر کرد. ویژگی LN این مهم را درنظر می‌گیرد.

ویژگی F مزایای ویژگی‌های NF و LN را با ترکیب این دو ویژگی استفاده می‌کند. بنابراین در کاربردهای خوشه‌بندی مقالات و یا طبقه‌بندی مقالات به موضوعات از پیش‌تعیین شده بهتر است که از این ویژگی استفاده گردد.

نتیجه‌گیری و پیشنهادات

نظر به اهمیت کلیدواژه‌ها در مقالات علمی و کاربردهای آن در حوزه‌های مختلف، این مقاله به بررسی ویژگی‌های مختلفی که می‌توان از آنها استخراج کرد، پرداخته است. علاوه بر آن، با استفاده از برچسب‌دهی که به کلیدواژه‌ها بر اساس هر کدام از ویژگی‌ها می‌دهد، می‌توان حوزه موضوعی یک متن را نیز استخراج نمود. از هر کدام از این ویژگی‌ها می‌توان در کاربردهای مختلف مانند ایجاد یک هستان‌شناسی از ارتباط میان کلیدواژه‌ها، ایجاد طبقه‌بند مقالات و یا خوشه‌بندی مقالات استفاده نمود. علاوه بر این، با استفاده از برچسب‌دهی موضوعی که از ویژگی‌های مبتنی بر کلیدواژه‌ها انجام می‌گردد، می‌توان حوزه موضوعی مقاله نیز مشخص کرد که این امر می‌تواند در بازیابی و باالاخص افزایش بازیافت سامانه‌های بازیابی اطلاعات استفاده نمود. به عنوان کارهای آینده، نویسنده برچسب‌زنی موضوعی بر اساس سطح‌بندی اسکوپوس و ارتباط این ویژگی‌ها با آن سرعنوان‌های موضوعی را بررسی می‌کند و همچنین ارتباط این ویژگی‌ها بر اساس کلیدواژه‌های همسایه در متن با استفاده از نظریه گرانش پیشنهاد می‌شود.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] گیلوری، عباس (۱۳۷۹). نمایه‌سازی خودکار (گذشته، حال، آینده). تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی (پیام کتابخانه سابق)، زمستان ۱۳۷۹. شماره ۳۹، ص ۱۷-۲۵.

[۲] آقابخشی، علی‌اکبر (۱۳۸۶). نمایه‌سازی همارا: مفاهیم و روش‌ها. تهران: چاپار: ۱۳۸۶.

[3] Lahr, N.B. and Barr, G.C., Synergy Sports Tech LLC, (2020). System and methods for searching and displaying ontology-based data structures. U.S. Patent Application 16/569,260.

[4] Dutta, S., Ghatak, S., Das, A.K., Gupta, M. and Dasgupta, S., (2019). Feature selection-based clustering on micro-blogging data. In Computational Intelligence in Data Mining (pp. 885-895). Springer, Singapore.

[5] Deng, X., Li, Y., Weng, J. and Zhang, J., (2019). Feature selection for text classification: A review. Multimedia Tools and Applications, 78(3), pp.3797-3816.

[6] Hitzler, P., Kirrane, S., Hartig, O., de Boer, V., Vidal, M.E., Maleshkova, M., Schlobach, S., Hammar, K., Lasierra, N., Stadtmüller, S. and Hose, K., (2019). The Semantic Web: ESWC 2019 Satellite Events. In ESWC2019, the Extended Semantic Web Conference (Vol. 11762). Springer.

[7] Ruotsalo, T. and Hyvönen, E., (2007). A method for determining ontology-based semantic relevance. In International Conference on Database and Expert Systems Applications (pp. 680-688). Springer, Berlin, Heidelberg.

[8] Ibrahim, R., Zeebaree, S. and Jacksi, K., (2019). Survey on Semantic Similarity Based on Document Clustering. Adv. Sci. Technol. Eng. Syst. J, 4(5), pp.115-122.

[9] Taloba, A.I., Sewisy, A.A. and Ismail, S.S., (2019). Parameter Tuning in Decision Tree Based on Genetic Algorithm for Text Classification. International Journal of Scientific & Engineering Resaerch, 10.

[10] Haddoud, M., Mokhtari, A., Lecroq, T. and Abdeddaïm, S., (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. Knowledge and Information Systems, 49(3), pp.909-931.

[11] Zuo, Z., Li, J., Anderson, P., Yang, L. and Naik, N., (2018), July. Grooming detection using fuzzy-rough feature selection and text classification. In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE.

[۱۲] ارسطوپور، شعله؛ آزاد، اسداله (۱۳۸۶). نظریه برچسب گذاری و برچسبهای موضوعی در سازماندهی اطلاعات: نگاهی تطبیقی از زاوه ارتباطهای متقاعدگرایانه. کتابداری و اطلاع رسانی. ۱۰ (۴): ۶۵-۸۸.

[13] Yang, Y. and Liu, X. (1999). A Re-examination of Text Categorization Methods, Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp. 42-49.

[14] Hiemstra, D. (2000). A probabilistic justification for using tf× idf term weighting in information retrieval. International Journal on Digital Libraries, 3(2), 131-139.

[15] Wood, S.A. and Gedeon, T.D. (2001). A Hybrid Neural Network for Automated Classification, Proceedings of the 6th Australasian Document Computing Symposium,2001.

[16] Torkolla, K. (2001). Linear Discriminant Analysis in Document Classification, *IEEE ICDM workshop on text mining*.

[17] Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993-1022.

[18] Guandong, X., Zhang, Y., Zhou, Z. (2005). Using Probabilistic Latent Semantic Analysis for Web Page Grouping, *Proceedings of Research Issues in Data Engineering: Stream Data Mining and Applications*, pp. 29-36.

[19] Jiang, S., Pang, G. Wu, M. and Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications* 39: 1503–1509.

[۲۰] عرب سرخی، محسن؛ فیلی، هشام (۱۳۸۵). ارائه یک سیستم دسته بندی موضوعی متون فارسی بر اساس روشهای احتمالاتی، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، صص ۱۵۱-۱۶۱.

ISSN: 2745-3673
http://www.Rcsj.ir

[۲۱] حاجی حسینی، آزاده؛ الماس گنج، فرشاد (۱۳۸۵). دسته بندی موضوعی متون فارسی بر اساس روش آنالیز معنایی پنهان، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، صص ۱۹۰-۲۰۱.

[22] Farhoodi, M., Yari, A., Mahmoudi, M. (2009). "A Persian Web Page Classifier Applying a Combination of Content-Based and Context-Based Features", International Journal of Information Studies, Vol. 1, No. 4.

[23] Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., & Bevacqua, A. (2013). Probabilistic topic models for sequence data. Machine learning, 93(1), 5-29.

[24] Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological review, 114(2), 211.

[25] Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 697-702). IEEE.

[26] Yang, G., Wen, D., Chen, N. S., & Sutinen, E. (2015). A novel contextual topic model for multi-document summarization. Expert Systems with Applications, 42(3), 1340-1352.

[27] Jameel, S., Lam, W., & Bing, L. (2015). Supervised topic models with word order structure for document classification and retrieval learning. Information Retrieval Journal, 18(4), 283-330.