مجله مدیریت اطلاعات سلامت
سال پانزدهم/ شماره پنجم/ آذر و دی ۱۳۹۷

Published by *Vesnu Publications*
DOI: 10.22122/him.v15i5.3560

مطالعه میزان شباهت اصطلاحات عنوان، کلید واژه‌های نویسنده و موضوعات کنترل شده برای تعیین فیلد مناسب در تحلیل‌های موضوعی علم‌سنجی

فریده عصاره[۱]، محمد توکلی‌زاده راوری[۲]، زاهد بیگدلی[۱]، رقیه قضاوی[۳]

مقاله پژوهشی

چکیده

مقدمه: به منظور انجام تحلیل‌های موضوعی در حوزه علم‌سنجی، این سؤال مطرح می‌گردد که کدام یک از فیلدهای کتاب‌شناختی حاوی موضوعات مورد تحلیل قرار گیرد. پژوهش حاضر با هدف مقایسه فیلدهای موضوعی مدارک و تعیین فیلد و یا ترکیبی از فیلدهای کامل و مناسب به منظور انجام تحلیل‌های مذکور انجام شد.

روش بررسی: این مطالعه به روش توصیفی و تحلیل واژگانی و با رویکرد علم‌سنجی انجام گرفت. به منظور انجام تحقیق، تولیدات علمی حوزه طب عملکردی گوارش به عنوان نمونه از پایگاه Scopus استخراج و تحلیل‌های مورد نظر بر روی ۱۳۷۹۸ مقاله دارای سه فیلد عنوان، کلید واژه و موضوعات کنترل شده انجام شد. پس از یکسان‌سازی داده‌ها، خوشه‌بندی به روش K-Means صورت گرفت و با محاسبه شاخص دربردارندگی برای خوشه‌های ایجاد شده، مشابهت موضوعات بین سه فیلد مورد نظر مشخص گردید.

یافته‌ها: بین فیلدهای کلید واژه عنوان و کلید واژه نویسندگان مشابهت بالایی (۸۷/۷۱ و ۸۵/۷۱) مشاهده شد. همچنین، پایین بودن مقدار شاخص دربردارندگی فیلد عنوان و موضوعات کنترل شده (صفر) حاکی از آن بود که مشابهت کمی بین زبان کنترل شده و واژگان مورد استفاده توسط نویسندگان در عنوان وجود داشت و نویسندگان واژگان مرجع را در عنوان استفاده نمی‌کنند.

نتیجه‌گیری: در صورت استفاده از واژگان فیلد عنوان، نتایج تحلیل زبان طبیعی را نشان خواهد داد، اما در صورتی که هدف، دسته‌بندی موضوعات به صورت منسجم باشد، استفاده از فیلد موضوعات کنترل شده، مناسب‌ترین فیلد خواهد بود.

واژه‌های کلیدی: تحلیل موضوعی؛ علم‌سنجی؛ عنوان؛ واژگان کنترل شده؛ کلید واژه‌ها

دریافت مقاله: ۱۳۹۷/۶/۲۳ پذیرش مقاله: ۱۳۹۷/۸/۲۸ تاریخ انتشار: ۱۳۹۷/۹/۱۵

ارجاع: عصاره فریده، توکلی‌زاده راوری محمد، بیگدلی زاهد، قضاوی رقیه. **مطالعه میزان شباهت اصطلاحات عنوان، کلید واژه‌های نویسنده و موضوعات کنترل شده برای تعیین فیلد مناسب در تحلیل‌های موضوعی علم‌سنجی.** مدیریت اطلاعات سلامت ۱۳۹۷؛ ۱۵ (۵): ۲۲۵-۲۲۰

مقدمه

رکوردهای موجود در پایگاه‌های اطلاعاتی دارای فیلدهای متفاوتی است که هر کدام حاوی یک فقره از اطلاعات کتاب‌شناختی مدارک می‌باشد (۱). بعضی از این فیلدها با هدف بازیابی مدارک و جهت استفاده کاربران نهایی و بعضی دیگر به منظور ارزیابی مدارک و استخراج آمار و استفاده سطح بالاتر کاربران مانند علم‌سنجان و سیاست‌گذاران علمی و پژوهشی ایجاد شده‌اند. مهم‌ترین نیاز در هنگام مواجهه با اطلاعات کتاب‌شناختی، جمع‌آوری و تنظیم داده‌ها با توجه به اهداف و الزامات مختلف با تأکید بر ویژگی‌های مختلف استخراج شده از همان داده‌ها است (۲). بنابراین، کاربران سطح بالا، کلیه فیلدها را با کاربری متفاوت مورد استفاده قرار می‌دهند؛ به طوری که علاوه بر تحلیل فراوانی نویسندگان، مجلات، کشورها و وابستگی سازمانی، برای تحلیل میزان رشد تولیدات علمی، سال انتشار مدارک؛ برای تحلیل استنادی، فیلد منابع در رکوردها؛ برای تحلیل همکاری‌های علمی، فیلدهای نویسنده و وابستگی سازمانی نویسندگان و برای تحلیل موضوعی فیلدهای عنوان، چکیده، کلید واژه و موضوعات کنترل شده یا توصیفگرها مورد استفاده قرار می‌گیرند. موضوعات کنترل شده، یک لیست مشخص از اصطلاحات با یک معنی ثابت و غیر قابل تغییر برای نمایه‌سازی مدارک می‌باشد تا از پراکندگی موضوعات مرتبط تحت عنوان‌های مختلف جلوگیری شود (۳).

در تحلیل‌های علم‌سنجی که بر روی موضوع تولیدات علمی انجام می‌شود، لازم است یک و یا تلفیقی از چند فیلد در اطلاعات کتاب‌شناختی مدارک به منظور انجام تحلیل انتخاب شود. همان‌گونه که اشاره شد، فیلدهای عنوان، چکیده، کلید واژه و موضوعات کنترل شده یا توصیفگرها برای انجام تحلیل‌های موضوعی مورد استفاده قرار می‌گیرد. این امکان در نرم‌افزارهای علم‌سنجی ایجاد شده است که بتوان یکی از فیلدهای موضوعی را جهت انجام فرایند تحلیل و ترسیم نقشه علمی به کار گرفت.

مقاله حاصل رساله دکتری تخصصی می‌باشد که با حمایت دانشگاه شهید چمران اهواز انجام شده است.

۱- استاد، کتابداری و اطلاع‌رسانی، گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شهید چمران اهواز، اهواز، ایران
۲- دانشیار، کتابداری و اطلاع‌رسانی، گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم اجتماعی، دانشگاه یزد، یزد، ایران
۳- دانشجوی دکتری تخصصی، علم اطلاعات و دانش‌شناسی، گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شهید چمران اهواز، اهواز، ایران (نویسنده طرف مکاتبه)
Email: r.ghazavi2011@gmail.com

با توجه به تفاوت‌هایی که در اختصاص موضوعات از سوی نمایه‌سازان و کاربرد اصطلاحات توسط نویسندگان در عنوان و کلید واژه‌ها وجود دارد (۴)، ترکیب اصطلاحات موجود در هر یک از این فیلدها متفاوت خواهد بود. بنابراین، در تحلیل‌های موضوعی که علاوه بر بسامد اصطلاحات و موضوعات، ارتباط آن‌ها را با یکدیگر مورد بررسی قرار می‌دهد و بر اساس آن تحلیل‌هایی همچون هم‌رخدادی واژگان، روند موضوعات، خوشه‌بندی موضوعات و تحلیل ساختار آن شکل می‌گیرد، انتخاب هر یک از فیلدهای در بردارنده اصطلاحات یا موضوعات، نتایج متفاوتی را ارایه خواهد داد. بنابراین، در این دسته از تحلیل‌های موضوعی در علم‌سنجی، همواره برای پژوهشگران این حوزه این سؤال مطرح است که کدام یک از این فیلدهای کتاب‌شناختی و یا تلفیق کدام یک از فیلدها، بهترین و کامل‌ترین گزینه برای تحلیل خواهد بود و نتایج به دست آمده انطباق مناسب‌تر و کامل‌تری با موضوعات مورد بحث در مدارک خواهد داشت؟

مطالعاتی در رابطه با تطابق و مقایسه میزان پوشش فیلدهای کتاب‌شناختی حاوی موضوعات در پایگاه‌های مختلف و با اهداف متفاوت انجام شده است. تعدادی از این تحقیقات به منظور مقایسه فیلدهای موضوعی مختلف با هدف بازیابی اطلاعات صورت گرفته است و گروهی دیگر از پژوهش‌ها به مقایسه کلید واژه‌های موجود در مدارک با اصطلاح‌نامه‌های مرتبط با یک حوزه موضوعی خاص و یا مقایسه پایگاه‌های اطلاعاتی مختلف در نمایه‌سازی موضوعی مدارک پرداخته‌اند.

Qin در مطالعه خود به دنبال کشف شباهت‌های اصطلاحات نمایه شده بین دو پایگاه اطلاعاتی Science Citation Index (SCI) و Medline بود. در تحقیق وی الگوهای توزیع فراوانی واژگان، شباهت بین واژگان برجسته، واژگان نسبتاً مشابه و خوشه‌های واژگان غیر مشابه تحلیل گردید (۴). Murphy و همکاران پژوهشی را به منظور مقایسه رفتار نویسندگان، جستجوگران و نمایه‌سازان در کاربرد واژگان انجام دادند. به این منظور، فراوانی سرعنوان‌های موضوعی پزشکی MeSH (Medical Subject Headings)، توصیفگرها و کلید واژه‌های مورد استفاده توسط نویسندگان در عناوین و چکیده‌ها در مقایسه با شیوه‌های استاندارد نمایه‌سازی معنایی و تحلیلی محاسبه شد (۵). مطالعه Gross با هدف ارزیابی تأثیر وجود و یا حذف فیلد سرعنوان موضوعی بر بازیابی مدارک انجام شد (۶).

در تحقیقی که توسط جوکار و انواری صورت گرفت، کارایی دو رویکرد جستجو با استفاده از زبان طبیعی و واژگان کنترل شده در دو پایگاه کتاب‌شناختی کتابخانه کنگره و Education Resources Information Center (ERIC) مقایسه گردید (۷). نتایج پژوهش نقنه اصفهانی و همکاران که با هدف میزان تطابق کلید واژه‌های فارسی و انگلیسی در عنوان و چکیده‌های پایان‌نامه‌های دانشگاه علوم پزشکی اصفهان با اصطلاح‌نامه پزشکی فارسی و اصطلاح‌نامه MeSH انجام شد، نشان داد که بین آن‌ها رابطه وجود دارد (۸). قنواتی و همکاران در مطالعه خود به بررسی تطابق و همخوانی سه نوع ساختار زبانی مهار شده اصطلاح‌نامه محور، مهار نشده کاربرمدار و کلید واژه‌های تخصیص داده شده به مدارک توسط نویسندگان در پایگاه اطلاعاتی ERIC و وبگاه Mendeley پرداختند (۹).

بر اساس بررسی‌های صورت گرفته بر روی تحقیقات مشابه در داخل و خارج از کشور که به مواردی از آن اشاره گردید، رویکرد غالب در این پژوهش‌ها، ذخیره، جستجو و بازیابی اطلاعات است و مطالعه‌ای که به مقایسه فیلدهای حاوی موضوعات و بررسی محدودیت‌های به کارگیری آن در مطالعات داده‌کاوی و علم‌سنجی پرداخته باشد، انجام نشده است. همچنین، در حوزه طب عملکردی گوارش نیز تحقیق مشابهی وجود ندارد. بنابراین، پژوهش حاضر با هدف مقایسه فیلدهای موضوعی مدارک و تعیین فیلد و یا ترکیبی از فیلدهای کامل و مناسب به منظور انجام تحلیل‌های موضوعی بر روی تولیدات علمی انجام شد. تبیین این موضوع به پژوهشگران در انتخاب فیلد مناسب برای انجام تحلیل‌های موضوعی در علم‌سنجی، بر اساس هدف در نظر گرفته شده برای هر مطالعه، کمک خواهد کرد.

روش بررسی

این مطالعه به روش توصیفی و تحلیل واژگانی و با رویکرد علم‌سنجی انجام شد. با توجه به محدودیت انجام تحقیق در کل حوزه‌های علمی، لازم بود تا تولیدات علمی یک حوزه علمی به عنوان نمونه انتخاب گردد. بر اساس مصاحبه با متخصصان حوزه‌های مختلف، مجموعه تولیدات علمی حوزه اختلالات عملکردی گوارش، با توجه به ویژگی‌هایی از جمله مناسب بودن کمیت تعداد مدارک در پایگاه‌های اطلاعاتی و تنوع کافی در کاربرد مفاهیم در بین متخصصان آن حوزه، جهت انجام پژوهش انتخاب گردید. همچنین، به دلیل نیاز به تحلیل سه فیلد عنوان، کلید واژه و موضوعات کنترل شده در رکوردهای هر مدرک، پایگاه اطلاعاتی Scopus به عنوان منبع استخراج مدارک برگزیده شد. استراتژی جستجو با به کار بردن مترادف‌های عبارات مربوط به این حوزه شامل «Functional Gastrointestinal Disorders ،Functional Diarrhea ،Functional Dyspepsia ،Functional Constipation ،Functional Bloating و Irritable Bowel Syndrome» تنظیم شد. پس از جستجو در فیلد عنوان و کلید واژه، در مجموع ۲۹۶۷۲ مقاله (تا پایان سال ۲۰۱۶) بازیابی گردید که فقط ۱۳۷۹۸ رکورد دارای هر سه فیلد مورد نظر (یعنی عنوان، کلید واژه و موضوعات کنترل شده) بودند. به منظور یکسان‌سازی فیلدهای مورد نظر، عملیاتی به صورت مشترک بر روی سه فیلد انجام گرفت.

مرحله اول، تبدیل کلمات جمع به مفرد بود که با استفاده از قواعد موجود در زبان انگلیسی به منظور جمع کردن کلمات (Plurals) و نیز با بررسی اصول مشترک در سایر کلمات جمع، تبدیل صورت گرفت. در مرحله دوم، کلماتی که دارای پسوند ing یا ed بود، به طور مشابه با مورد قبل و بر اساس دستور زبان انگلیسی، به کلمات ساده تبدیل گردید. در مرحله سوم، لیست Stop Words در فیلد عنوان به منظور نمایه‌سازی این فیلد و تبدیل جمله به عبارات موضوعی، با توجه به این که این کلمات محل‌هایی به منظور شکاف جمله می‌تواند در نظر گرفته شود، به علامت جدا کننده (که با توجه به ویژگی منحصر به فرد بودن علامت # در نظر گرفته شده است) تبدیل شد. در مرحله چهارم، علایم نگارشی مورد استفاده در فیلد عنوان به علامت # تغییر یافت. در مرحله پنجم، بر اساس روش شکاف و گلچین (۱۰)، تک‌واژه‌های کلیه کلمات و عبارات موضوعی به دست آمده در هر یک از سه فیلد به طور مجزا استخراج گردید. در نهایت، ترکیب عبارات موضوعی و تک‌واژه‌های استخراج شده از آن در تحلیل‌های بعدی وارد شد. در مرحله ششم و به منظور مقایسه موضوعات هر یک از سه فیلد با دو فیلد دیگر، لازم بود دسته‌بندی مناسب بر آن‌ها انجام شود و دسته‌های حاصل با یکدیگر مقایسه گردد. با توجه به این ویژگی، روش خوشه‌بندی

K-Means انتخاب شد. روش مذکور این امکان را فراهم می‌کند تا تعداد خوشه‌ها توسط کاربر و با توجه به تناسب نتایج حاصل از خوشه‌بندی، تعیین گردد. این نوع خوشه‌بندی برای آماده‌سازی داده‌ها جهت اعمال خوشه‌بندی ماتریس تفاضل فراوانی به صورت موضوعات با استفاده از نرم‌افزار PreMap و نرم‌افزار تهیه شده به این منظور به زبان برنامه‌نویسی C به دست آمد. ماتریس حاصل در نرم‌افزار SPSS نسخه ۱۹ (version 19, SPSS Inc., Chicago, IL) وارد و خوشه‌بندی با تعداد دسته‌های مختلف بر این سه فیلد آزمون گردید. اجرای خوشه‌بندی با استفاده از تعداد خوشه‌های مختلف، به منظور بررسی توزیع مناسب موضوعات و رسیدن به یک سطح قابل قبول و قابل مدیریت (۱۱) بود. مقایسه نتایج حاصل از خوشه‌بندی با تعداد خوشه‌های ۳، ۴، ۵، ۷، ۱۰، ۱۵ و ۳۰ نشان داد که توزیع کلید واژه‌ها در ۷ خوشه تناسب بیشتری نسبت به سایر دسته‌ها دارد. در مرحله هفتم، میزان تشابه موضوعات خوشه‌های یک فیلد نسبت به خوشه‌های فیلدهای دیگر مورد مقایسه قرار گرفت. بدین منظور، از شاخص دربردارندگی با رابطه ۱ استفاده گردید (۱۳، ۱۲) که در آن، N معادل موضوعات مشترک در گروه A و B و M تعداد موضوعات قرار گرفته در گروه A می‌باشد.

$$IncA,B = 100\frac{N}{M}$$ رابطه ۱

این شاخص، دربردارندگی به صورت یک‌طرفه را نشان می‌دهد. دربردارندگی خوشه‌های بالاتر (فراوانی بیشتر) که واژگان پایه قلمداد می‌گردد (۱۴)، اهمیت بیشتری دارد. بنابراین، شاخص دربردارندگی در چندین مرحله محاسبه شد؛ به طوری که این شاخص در ابتدا در کل خوشه‌ها و سپس در ۴ خوشه اول، ۵ خوشه اول و ۶ خوشه اول محاسبه گردید. در مرحله هشتم، آزمون Friedman به منظور رتبه‌بندی هر یک از فیلدها بنا بر شاخص دربردارندگی به دست آمد.

یافته‌ها

یافته‌های حاصل از تعداد موضوعات قرار گرفته در هر یک از خوشه‌ها در فیلدهای مختلف در جدول ۱ ارایه شده است.

جدول ۱: تعداد موضوعات موجود در هر یک از خوشه‌ها در سه فیلد موضوعی

خوشه	کلید واژه نویسندگان	کلید واژه عنوان	موضوعات کنترل شده
۱	۱	۱	۱
۲	۱	۱	۱
۳	۲	۲	۲
۴	۳	۳	۷
۵	۱۵	۱۷	۱۶
۶	۱۰۵	۱۱۳	۲۰۰
۷	۸۶۹۳	۷۲۲۱	۱۵۷۴۸
جمع	۸۸۲۰	۷۳۵۸	۱۵۹۷۵

با استفاده از شاخص دربردارندگی، کلید واژه‌های قرار گرفته در هر یک از خوشه‌ها برای سه فیلد به صورت دو به دو مقایسه گردید و نتایج آن در جدول مقایسات زوجی (جدول ۲) وارد شد. بر این اساس، بالاترین میزان دربردارندگی به ترتیب با مقادیر ۸۷/۷۱ و ۸۵/۷۱ مربوط به کلید واژه‌های عنوان و کلید واژه‌های نویسندگان بود.

جدول ۲: شاخص دربردارندگی برای سه فیلد کلید واژه نویسندگان، کلید واژه عنوان و موضوعات کنترل شده به صورت دو به دو در تقسیم‌هایی از ۷ خوشه

کلید واژه	کل خوشه‌ها (۷ خوشه)	٦ خوشه اول	٥ خوشه اول	٤ خوشه اول
کلید واژه عنوان– کلید واژه نویسنده	۳۶/۹۶	۶۷/۷۱	۷۲/۷۲	۸۷/۷۱
کلید واژه عنوان– موضوعات کنترل شده	۲۸/۵۷	۵۳/۵۴	۳۱/۸۱	۰
کلید واژه نویسنده– موضوعات کنترل شده	۴۳/۳۴	۶۲/۰۴	۴۵/۸۳	۱۴/۲۸
کلید واژه نویسنده– کلید واژه عنوان	۴۴/۳۰	۶۲/۷۷	۶۶/۶۶	۸۵/۷۱
موضوعات کنترل‌شده– کلید واژه عنوان	۱۵/۷۷	۲۹/۹۵	۲۵/۹۲	۰
موضوعات کنترل‌شده– کلید واژه نویسنده	۱۹/۹۶	۳۷/۴۴	۴۰/۷۴	۹/۰۹

داده‌های حاصل از آزمون Friedman به منظور تعیین رتبه هر یک از فیلدهای موضوعی بر اساس شاخص دربردارندگی به صورت یک‌طرفه و درهمکرد دو طرفه به ترتیب در جداول ۳ و ۴ ارایه شده است.

جدول ۳: نتایج تحلیل آماری با استفاده از روش Friedman و تعیین رتبه هر یک از فیلدها (درهمکرد دو طرفه)

نوع کلید واژه	رتبه	میانگین ± انحراف معیار	حد پایین	حد بالا
کلید واژه نویسنده	۲/۶۳	۵۳/۱۲ ± ۲۱/۲۲	۱۴/۲۸	۸۵/۷۱
کلید واژه عنوان	۲/۱۳	۴۷/۳۸ ± ۲۸/۴۶	۰	۸۷/۷۱
موضوعات کنترل شده	۱/۲۵	۲۲/۳۶ ± ۱۳/۹۴	۰	۴۰/۷۴

مطابق با درصدهای دربردارندگی، P = ۰/۰۲۱ و χ^2 = ۷/۷۵، بالاترین رتبه میزان دربردارندگی (۲/۶۳) به کلید واژه نویسندگان اختصاص داشت (جدول ۳).

بنا بر یافته‌های ارایه شده در جدول ۴ و بر اساس درصدهای دربردارندگی، P = ۰/۰۰۳ و χ^2 = ۱۷/۸۷، بالاترین رتبه میزان دربردارندگی یک‌طرفه (۵/۵) مربوط به کلید واژه عنوان در کلید واژه نویسندگان بود.

جدول ۴: نتایج تحلیل آماری با استفاده از روش Friedman و رتبه برای هر یک از فیلدها (به صورت یک‌طرفه)

نوع کلید واژه	رتبه	میانگین ± انحراف معیار	حد پایین	حد بالا
کلید واژه عنوان- کلید واژه نویسنده	۵/۵	۶۶/۲۸ ± ۲۱/۳۱	۳۶/۹۶	۸۷/۷۱
کلید واژه نویسنده- کلید واژه عنوان	۵/۲۵	۶۴/۸۶ ± ۱۶/۹۸	۴۴/۳۰	۸۵/۷۱
کلید واژه نویسنده- موضوعات کنترل شده	۴/۲۵	۴۱/۳۷ ± ۱۹/۸۷	۱۴/۲۸	۶۲/۰۴
موضوعات کنترل شده- کلید واژه نویسنده	۲/۵	۲۶/۸۱ ± ۱۴/۹۲	۹/۰۹	۴۰/۷۴
کلید واژه عنوان- موضوعات کنترل شده	۲/۳۸	۲۸/۴۸ ± ۲۱/۹۹	۰	۵۳/۵۴
موضوعات کنترل شده- کلید واژه عنوان	۱/۱۳	۱۷/۹۱ ± ۱۳/۳۵	۰	۲۹/۹۵

بحث

با توجه به این که خوشه‌بندی به روش K-Means انجام شد، اساس دسته‌بندی، بسامد موضوعات می‌باشد. بنابراین، تعداد موضوعات در خوشه‌های اول محدود است و تنها مواردی با فراوانی بالاتر قرار دارند و بر عکس در خوشه‌های انتهایی تعداد موضوعات زیاد، اما با بسامد کم جای می‌گیرند. این دسته‌بندی در پژوهش‌هایی که به روش Bradford انجام شده است، نیز مشاهده می‌شود (۱۲).

طبق میزان دربردارندگی یک‌طرفه موضوعات به صورت دو به دو، نتایج بر اساس موضوعات برجسته (موضوعات مربوط به خوشه‌های اول) در دسته‌بندی‌های مختلف نشان داد که دربردارندگی بین کلید واژه عنوان و نویسنده به صورت دو طرفه، بالاترین میزان را دارد. بعد از دربردارندگی این دو فیلد، فراوانی یک‌طرفه کلید واژه نویسنده در موضوعات کنترل شده، بالاترین میزان را به خود اختصاص داد؛ چرا که در علوم پزشکی نویسندگان موظف به استفاده از اصطلاحنامه MeSH در انتخاب کلید واژه برای مقالات هستند. بین موضوعات کنترل شده و کلید واژه عنوان به صورت دو طرفه در همه دسته‌بندی‌ها، پایین‌ترین میزان دربردارندگی مشاهده گردید. کلید واژه‌های نویسندگان نیز حالت بینابینی دارد؛ به طوری که تعداد بالاتری از کلید واژه‌های نویسندگان به نسبت کلید واژه‌های عنوان، در موضوعات کنترل شده وجود داشت. این یافته همان‌گونه که در پژوهش‌های حوزه بازیابی اطلاعات نیز مشخص شده است، حاکی از کاربرد اصطلاحات متفاوت از موضوعات کنترل شده توسط نویسندگان در عنوان و حتی کلید واژه‌ها می‌باشد. مقایسه بین کلمات عناوین و کلمات اصطلاحات کنترل شده اختصاص یافته در مطالعه جوکار و انواری نیز نشان داد که کلمات عناوین درصد کمی از مفاهیم موجود در موضوعات را تحت پوشش قرار می‌دهند (۷). همچنین، تحقیق قنواتی و همکاران به این نتیجه رسیدند که برچسب‌گذاران واژه‌هایی متفاوت از توصیفگرهای نمایه‌سازان و کلید واژه‌های نویسندگان استفاده کرده‌اند که نشان دهنده عدم آشنایی سه گروه به زبان و واژگان مورد استفاده یکدیگر است (۹). در بازیابی مدارک نیز Murphy و همکاران دریافتند که به دلیل عدم تطابق بین موضوعات کنترل شده و فرایند نمایه‌سازی پایگاه‌های اطلاعاتی، نیاز به یک واژه‌شناختی استاندارد بین جستجوگران در تنظیم استراتژی جستجو و نویسندگان در فرایند نوشتن عنوان، چکیده و کلید واژه‌ها وجود دارد (۵).

بر اساس رتبه‌بندی مقادیر دربردارندگی دو طرفه هر یک از فیلدها، کلید واژه نویسنده بالاترین میزان دربردارندگی را نسبت به دو فیلد دیگر دارد که در تحلیل دربردارندگی یک‌طرفه، به طور خاص پوشش کلید واژه عنوان توسط کلید واژه‌های نویسنده بالاترین رتبه را به خود اختصاص داد. مطابق با نتایج به دست آمده، پایین‌ترین رتبه متعلق به پوشش موضوعات کنترل شده توسط کلید واژه‌های عنوان می‌باشد. در این رابطه نیز در پژوهش‌های حوزه بازیابی اطلاعات مانند مطالعه Gross (۶)، اهمیت موضوعات کنترل شده در بازیابی مشخص گردید؛ به طوری که در بیش از یک سوم مدارک بازیابی شده با جستجوی کلید واژه‌ای، اگر سرعنوان‌های موضوعی وجود نداشتند، از دست می‌رفتند (۶) که با نتایج تحقیق حاضر همخوانی نداشت.

نتیجه‌گیری

بر اساس موارد ذکر شده در پژوهش حاضر، چنین می‌توان استنتاج کرد که با در نظر گرفتن و تأکید بر هدف مورد نظر در تحلیل‌های علم‌سنجی، می‌توان تحلیل را با استفاده از یکی از فیلدهای سه‌گانه اجرا نمود. با توجه به بالا بودن شاخص دربردارندگی دو طرفه فیلدهای کلید واژه عنوان و کلید واژه نویسندگان، می‌توان در تحلیل‌های علم‌سنجی از هر یک از این فیلدها به جای فیلد دیگر استفاده کرد و نتایج به نسبت مشابهی را انتظار داشت. همچنین، با توجه به پایین بودن مقدار شاخص دربردارندگی دو طرفه فیلدهای کلید واژه عنوان و موضوعات در صورت استفاده از کلید واژه‌های عنوان، نتایج تحلیل زبان طبیعی را نشان خواهد داد. بنابراین، در مواردی که استفاده از زبان طبیعی نویسندگان در به کار بردن واژگان مربوط به یک حوزه و سیر تغییرات آن مد نظر باشد، فیلدهای عنوان و کلید واژه نویسندگان می‌تواند استفاده گردد، اما فیلد عنوان مناسب‌تر می‌باشد. چنانکه Davis نیز اشاره می‌کند که استفاده از فیلد عنوان مقاله، اولین رویکرد در استفاده از زبان طبیعی نویسندگان خواهد بود (۱۵). به خصوص در مدارکی که به سال‌های قبل از ۲۰۰۰ تعلق دارد و کلید واژه نویسندگان در پایگاه‌های اطلاعاتی ارایه نشده است، نمایه‌سازی عنوان می‌تواند جایگزین مناسبی برای تحلیل‌های موضوعی باشد؛ چرا که در بعضی از پژوهش‌های علم‌سنجی، نویسندگان مجبور به حذف مدارک فاقد کلید واژه نویسندگان می‌شوند (۱۶). همچنین، اگر هدف نویسنده، دسته‌بندی موضوعات به صورت منسجم و به دور از پراکندگی استفاده از واژگان به صورت ناهمگون باشد، استفاده از فیلد موضوعات کنترل شده، مناسب‌ترین فیلد خواهد بود.

نتایج مطالعه حاضر که به طور نمونه بر روی حوزه طب عملکردی گوارش انجام شد، می‌تواند در سایر تحقیقات علم‌سنجی بر موضوعات دیگر نیز به کار برده شود. این نکته به خصوص در رابطه با پژوهش‌هایی که موضوع انتخاب شده جهت تحلیل در حوزه علوم پزشکی و پیراپزشکی می‌باشد، صادق است؛ چرا

که در این حوزه‌ها استفاده از اصطلاح‌نامه MeSH، مرجع انتخاب کلید واژه نویسندگان است و بیشتر مجلات این حوزه‌ها در این رابطه از الگوی یکسانی پیروی می‌کنند. همچنین، در پایگاه Scopus نیز مرجع انتخاب موضوعات کنترل شده به طور عمده اصطلاح‌نامه MeSH می‌باشد.

از جمله محدودیت‌های مطالعه حاضر می‌توان به عدم امکان استفاده از مقالات مربوط به موضوع طب عملکردی گوارش در پایگاه Web of Science و PubMed اشاره کرد که با توجه به تفاوت نمایه‌سازی این دو پایگاه و در نتیجه، اختصاص موضوعات کنترل شده به مقالات به صورت متفاوت از پایگاه اطلاعاتی مورد استفاده در این تحقیق، امکان‌پذیر نبود. همچنین، به منظور عدم وجود ابزار بهینه‌ای جهت پردازش زبان طبیعی و نمایه‌سازی خودکار فیلدهای حاوی موضوعات، انجام مراحل لازم به منظور پردازش مذکور به صورت دستی و با صرف زمان زیاد، در پژوهش حاضر انجام پذیرفت.

پیشنهادها

پیشنهاد می‌گردد که در بخش روش‌شناختی پژوهش‌های علم‌سنجی، علاوه بر اشاره به فیلد مورد جستجو جهت بازیابی مدارک پایه، به فیلد موضوعی مورد تحلیل نیز اشاره گردد. همچنین، بهتر است محققان قبل از انتخاب هر یک از فیلدهای موضوعی برای انجام تحلیل‌های علم‌سنجی، تحلیل و بررسی اولیه‌ای بر مفاهیم و موضوعات ارایه شده انجام دهند تا ضمن مشورت با متخصصان موضوعی مربوط، نسبت به تفاوت برون‌داد هر یک از فیلدهای مذکور شناخت نسبی پیدا کنند.

تشکر و قدردانی

بدین وسیله از راهنمایی‌های جناب آقای دکتر پیمان ادیبی، فوق تخصص بیماری‌های گوارش و کبد و عضو محترم هیأت علمی دانشگاه علوم پزشکی اصفهان تشکر و قدردانی به عمل می‌آید.

تضاد منافع

در انجام مطالعه حاضر، نویسندگان هیچ‌گونه تضاد منافعی نداشته‌اند.

References

1. Jafarnejad A. An introduction to data banks. Tehran, Iran: SAMT Publications; 2006. p. 24. [In Persian].
2. Ferrara A, Salini S. Ten challenges in modeling bibliographic data for bibliometric analysis. Scientometrics 2012; 93(3): 765-85.
3. Prytherch RJ, Harrod LM. Harrod's Librarians' glossary of terms used in librarianship, documentation and the book crafts, and reference book. Aldershot, UK: Gower; 1990. p. 163.
4. Qin J. Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. J Am Soc Inf Sci 2000; 51(2): 166-80.
5. Murphy LS, Reinsch S, Najm WI, Dickerson VM, Seffinger MA, Adams A, et al. Searching biomedical databases on complementary medicine: The use of controlled vocabulary among authors, indexers and investigators. BMC Complement Altern Med 2003; 3: 3.
6. Gross Ti. What have we got to lose? The effect of controlled vocabulary on keyword searching results. College and Research Libraries 2018; 66(3): 212-30.
7. Jokar A, Anvari S. Study of thematic approaches (Natural and controlled language) in information retrieval from online bibliographic databases. Library and Information Science 2007; 9(4): 151-64. [In Persian].
8. Naghneh Esfahani M, Cheshmeh Sohrabi M, Banieghbal N. A comparative study of the Persian and English keywords of theses from the isfahan university of medical sciences, Iran, and the thesauruses and Persian medical subject headings. Health Inf Manage 2013; 9(6): 802-13. [In Persian].
9. Ghanavati M, Noruzi A, Nakhoda M, Khatir A. Consistency between descriptors, author-supported keywords and tags in the ERIC and Mendeley databases. Iranian Iranian Journal of Information Processing and Management 2018; 33(4): 1745-66. [In Persian].
10. Tavakolizadeh-Ravari M. Two steps break-cull model for automatic indexing of Persian texts. Research on Information Science and Public Libraries 2015; 21(80): 13-40. [In Persian].
11. Tseng YH, Tsay MY. Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. Scientometrics 2013; 95(2): 503-28.
12. Mokhtari-Shamsi M, Tavakolizadeh-Ravari M, Zalzadeh E, Baghbanian M. Predicting basic concepts of a field, based on the factors of oldness and frequency use of subject terms: A case study on colon cancer. Health Inf Manage 2016; 13(5): 354-9. [In Persian].
13. Hazeri A, Tavakolizadeh Ravari M, Ebrahimi V. A study of subject overlap between the main categories of knowledge management within the web of science. Iranian Journal of Information Processing and Management 2015; 30(4): 997-1023. [In Persian].
14. Hosaininasab SH, Makkizadeh F, Zalzadeh E, Hazeri A. The thematic structure of papers on depression treatment in PubMed from 2005 to 2014. Health Inf Manage 2016; 13(5): 347-53. [In Persian].
15. Davis MA. Title keyword selection and use for optimum document retrieval. Public and Access Services Quarterly 1997; 2(2): 15-22.
16. Makizadeh F. The semantic relationship between the themes in Persian scientific articles in the field of global warming. Journal of Climate Research 2016; 7(25-26): 91-109. [In Persian].

Health Information Management

Published by ***Vesnu Publications***

Vol 15/ No 5/ December 2018

DOI: 10.22122/him.v15i5.3560

Study of Similarities of Terms in Title, Author's Keywords and Controlled Vocabulary for Determining the Appropriate Field in Scientometric Thematic Analysis

Farideh Osareh[1], Mohammad Tavakolizadeh-Ravari[2], Zahed Bigdeli[1], Roghayeh Ghazavi[3]

Original Article

Abstract

Introduction: One problem in conducting scientometric thematic analysis is selecting which of the bibliographic fields containing the topics can be analyzed. This study aims to compare subject fields of documents to determine the field or a combination of fields which are suitable for conducting a complete and proper thematic analysis in scientometrics.

Methods: This was a descriptive research with content analysis approach. The scientific products in the field of functional gastrointestinal disorders were extracted from the Scopus database. The analysis was done on 13798 documents, which included title, author keywords and index keywords. After clustering using the K-Means method and calculating the inclusion index for the created clusters, the similarity of the keywords between the three fields was determined.

Results: The results showed that there is a high similarity between the index keywords and the author keywords (87.71 and 85.71). The low amount of the index in the title field and the index keywords (0) also suggests that there is little similarity between the controlled Vocabulary and the keywords used by the authors in the title, and that authors do not use the preferred vocabulary in the title.

Conclusion: Using the words of the title field will show the results of the natural language analysis. However, if the purpose of a study is categorizing terms, the use of index keywords field will be the most appropriate.

Keywords: Thematic Analysis; Scientometrics; Title; Controlled Vocabulary; Keywords

Received: 14 Sep., 2018 Accepted: 19 Nov., 2018 Published: 06 Dec., 2018

Citation: Osareh F, Tavakolizadeh-Ravari M, Bigdeli Z, Ghazavi R. **Study of Similarities of Terms in Title, Author's Keywords and Controlled Vocabulary for Determining the Appropriate Field in Scientometric Thematic Analysis.** Health Inf Manage 2018; 15(5): 220-5

Article resulted from PhD thesis funded by Shahid Chamran University of Ahvaz.

1- Professor, Knowledge and Information Science, Department of Knowledge and Information Science, School of Educational Sciences and Psychology, Shahid Chamran University of Ahvaz, Ahvaz, Iran

2- Associate Professor, Knowledge and Information Science, Department of Knowledge and Information Science, School of Social Sciences, Yazd University, Yazd, Iran

3- PhD Student, Knowledge and Information Science, Department of Knowledge and Information Science, School of Educational Sciences and Psychology, Shahid Chamran University of Ahvaz, Ahvaz, Iran (Corresponding Author) Email: r.ghazavi2011@gmail.com