تشخیص خودکار شاخه موضوعی اصطلاحات سرعنوان‌های موضوعی پزشکی با مقایسه نسبت فراوانی آن‌ها در مدارک مرتبط و غیرمرتبط*

محمد توکلی‌زاده راوری[1]، سعید غفاری[2]، فروغ مصطفوی[3]

مقاله پژوهشی

چکیده

مقدمه: تحت تاثیر پویایی اصطلاحات تخصصی، امروزه طبقه بندی موضوعات پیچیده تر شده است زیرا هر مدرک می تواند در چند طبقه موضوعی جای بگیرد. بر این اساس، پژوهش حاضر با هدف تعیین کارآمدی روش تشخیص خودکار شاخه اصلی اصطلاحات Medical Subject Heading از طریق محاسبه نسبت فراوانی آن‌ها در دسته مدارک مرتبط وغیرمرتبط انجام شد.

روش بررسی: روش پژوهش توصیفی، با استفاده از تحلیل اسنادی و نوع آن کاربردی است. در تیر ماه ۱۳۹۱ شمسی از MeSH و پایگاه PubMed به عنوان منابع گردآوری اطلاعات بهره گرفته شد. اعتبار این منابع، روا بودن بهره گیری از آن‌ها را تایید می کند. تعداد ۱۸۱۶۴ اصطلاح MeSH و ۱۶۳۲۲۶ مدرک از PubMed برگزیده شد. در گزینش آن‌ها، هیچ محدودیت زمانی اعمال نشد. این تعداد، از حجم نمونه به روش کوکران بالاتر بود. با جستجو در PubMed، یازده دسته مدرک حاصل شد. نسبت حضور هر اصطلاح در این دسته‌ها محاسبه و نتیجه با شاخه واقعی آن در درخت MeSH مقایسه شد. شاخه اصلی یک درصد از این اصطلاحات توسط متخصصان پزشکی نیز پیش‌بینی گردید. برای بررسی داده‌ها، از روش توزیع فراوانی و آزمون‌های T و Chi-Squar بهره گرفته شد. تحلیل داده‌ها با نرم‌افزار SPSS صورت گرفت.

یافته‌ها: مدارک PubMed به طور متوسط به سه شاخه مربوط بودند و غالب اصطلاحات در تمامی دسته‌ها حضور داشتند. مشخص شد که روش پیشنهادی، احتمال تشخیص منطبق با ساختار درخت موضوعی MeSH را افزایش می دهد و کارآمدی آن بسته به شاخه موضوعی، بین ۳ تا ۶۷ درصد متفاوت است. پیش‌بینی متخصصان پزشکی درباره شاخه موضوعی هر اصطلاح، به طور معناداری با ساختار MeSH منطبق بود.

نتیجه‌گیری: سطح انطباق تشخیص طبقه موضوعات به روش‌های عینی و ذهنی در حوزه های گوناگون فرق می کند. از آن جا که طبقه بندی‌های ذهنی کاری کاملا ادراکی و مربوط به تجربه‌های بیرونی بشری است، مدل های ماشینی نمی توانند دقیقا آن فرآیند را مشابه سازی کنند.

واژه‌های کلیدی: طبقه بندی؛ پردازش خودکار داده‌ها؛ سرعنوان‌های موضوعی پزشکی؛ PubMed

دریافت مقاله: ۹۲/۹/۱ اصلاح نهایی: ۹۳/۱/۲۳ پذیرش مقاله: ۹۳/۳/۱۱

ارجاع: توکلی‌زاده راوری، غفاری سعید، مصطفوی فروغ. **تشخیص خودکار شاخه موضوعی اصطلاحات سرعنوان‌های موضوعی پزشکی با مقایسه نسبت فراوانی آن‌ها در مدارک مرتبط وغیر مرتبط.** مدیریت اطلاعات سلامت ۱۳۹۴؛ ۱۲(۱): ۴۸-۶۰.

*- این مقاله حاصل یک پژوهش مستقل است.

۱- استادیار، علم اطلاعات و دانش شناسی، گروه علم اطلاعات و دانش‌شناسی، دانشگاه یزد، یزد، ایران(نویسنده مسؤول) Email: tavakoli@yazd.ac.ir

۲- استادیار، علم اطلاعات و دانش شناسی، گروه علم اطلاعات و دانش شناسی، دانشگاه پیام نور قم، قم، ایران

۳- کارشناس ارشد، علم اطلاعات و دانش شناسی، دانشگاه پیام نور اقلید، اقلید، ایران

مقدمه

مرزبندی میان رشته‌ها امروزه به سادگی امکان‌پذیر نیست (۱). Krauthammer و Nenadic بیان می‌دارند: «به علت تحولات پویا در زمینه اصطلاحات زیست پزشکی، موضوع تشخیص اصطلاح به عنوان یک مساله در داده کاوی متن در آمده است. در پی آن، این مبحث به عنوان یک موضوع پژوهشی مهم برای پردازش زبان و نهادهای زیست پزشکی تبدیل شده است» (۲).

تحت تاثیر در هم تنیده شدن مرزهای دانش و پویایی اصطلاحات تخصصی، امروزه طبقه‌بندی موضوعات پیچیده‌تر شده است زیرا هر مدرک می‌تواند در چند طبقه موضوعی جای بگیرد. همین‌طور، هر اصلاح موضوعی می‌تواند در مدارکی ظاهر شود که در عمل به یک یا چند حوزه موضوعی دیگر (غیر از طبقه از پیش تعیین شده) مربوط هستند. از طرفی حجم بالای اطلاعات باعث شده است که دسته‌بندی و طبقه‌بندی اطلاعات با فنون مختلف به صورت خودکار صورت گیرد. همه این مسائل منجر به ادبیاتی گردیده است که به تشخیص خودکار طبقه یا حوزه اصطلاحات موضوعی می‌پردازد. ادبیات این پژوهش‌ها با وجود شباهت، تفاوت‌هایی با پژوهش‌هایی دارد که به تشخیص خودکار اصطلاحات موضوعی یک مدرک و ارتباط سنجی آن اصطلاحات می‌پردازد. Krauthammer و Nenadic در بیان تفاوت تشخیص اصطلاح و تشخیص طبقه یک اصطلاح اظهار می‌دارند: «تشخیص اصطلاح روشی برای تعیین یک واحد زبانی است که به مفاهیم یک حوزه مربوط می‌شود. نقش طبقه‌بندی اصطلاح مشخص ساختن نوع خاصی از مفهوم یک حوزه(مانند ژن، پروتئین و mRNA) است که توسط اصطلاح توصیف شده است. به عبارت دیگر، طبقه‌بندی اصطلاحات، چراغ راهی اولیه برای تعیین اصطلاحات و به عنوان یک قدم مهم به سوی مرحله نهایی تشخیص اصطلاح محسوب می‌شود»(۲). این گفتار مطابق با گفته Vu و همکاران است که اصطلاحات موضوعی را سلسله‌ای از «واحدهای زبانی»(Unithood)می‌دانند که «محتوای یک مدرک» (Termhood) را توصیف می‌کنند. پس در عمل، تشخیص محتوای یک مدرک رابطه مستقیمی با تشخیص طبقه آن دارد (۳).

تمرکز پژوهشگران در این پژوهش بر تشخیص طبقه اصطلاحات به صورت نیمه عینی یا شبه نظارت شده است. این روش ترکیبی از تصمیم‌گیری توسط انسان و ماشین است. در این راستا، پرسش‌های زیر مورد توجه قرار خواهد گرفت:

۱- در عمل، هر مدرک از PUBMED را درچند طبقه موضوعیMeSH می‌توان جای داد؟

۲- در عمل، یک اصطلاح MeSH را در چند طبقه موضوعی می‌توان جای داد؟

۳- آیا روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط و غیر مرتبط، احتمال تشخیص خودکار شاخه اصلی اصطلاحات MeSH را به طور معنادارافزایش می‌دهد؟

۴- آیا عملکرد روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط وغیر مرتبط برای تشخیص خودکار شاخه اصلی اصطلاحات MeSH، در همه شاخه‌هایMeSH به یک اندازه است؟

۵- تخمین متخصصان پزشکی درباره طبقه موضوعی اصطلاحات، چقدر بر ساختار سرعنوان‌های موضوعیMeSH منطبق است؟

۶- تخمین متخصصان پزشکی درباره طبقه موضوعی اصطلاحات، چقدر بر تشخیص روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط وغیر مرتبط منطبق است؟

موضوع اصلی نوشتار حاضر، تعیین خودکار حوزه اصطلاحات موضوعی بر اساس بسامد اصطلاحات درمدارک مربوط و نامربوط است. از این رو، به ادبیاتی پرداخته می‌شود که به بسامد اصطلاحات در مدارک مربوط و نامربوط پرداخته است و پس از آن، به ادبیات تعیین و تشخیص طبقه، شاخه، حوزه یا رشته موضوعی توجه خواهد شد. شاید بتوان مرتبط‌ترین اثر به پژوهش حاضر را، اثر Harter دانست (۴). Harter در این راستا فرمولی را ارائه کرده است که در آن نسبت تعداد مدارکی که اصطلاح مورد نظر بسامد بیشتری دارد به تعداد کل مدارک موجود در آن حوزه سنجیده می‌شود:

$$\Pr\left(f(w_{ij})=x\right)=\pi\frac{e^{-m_{1i}}\times m_{1i}^{x}}{x!}+(1-\pi)\frac{e^{-m_{2i}}\times m_{2i}^{x}}{x!}$$

در فرمول بالا، $\Pr\left(f(w_{ij})=x\right)$ نسبت تعداد مدارکی که رخداد اصطلاح w_i در آن ها x بار است بـر تعـداد کـل مدارک می‌باشد. همچنین، m_{1i} و m_{2i} به ترتیـب میـانگین رخداد اصطلاح w_i در مـدارک گـروه ۱ (مـدارک مربـوط) و مدارک گروه ۲ (مـدارک نـامربوط) اسـت. نهایتـا، π نسـبت مدارک گروه اول به کل مدارک است.

بـر اسـاس گفتـه Kageura و mino (۵) آثـاری ماننـد Bookstein و Swanson(۶)، okstein وSwanson (۷)، Maron وCooper (۸) به موضوع وزن یـک اصـطلاح در مـدارک مربـوط و نـامربوط پرداختـه‌انـد کـه Salton (۹) اساس محاسبات کار آن‌ها را به صورت زیر بیان کرده است:

$$I_i=\log\frac{p_i(1-q_i)}{q_i(1-p_i)}$$

دربالا، p_i امکان رخداد اصطلاح مورد نظردر مدارک مربوط و q_i احتمال رخداد آن اصطلاح در مدارک نامربوط است. بایـد توجه داشت که آثار یاد شده در بالا، تلاششـان بـر تشـخیص اصطلاحات موضوعی مدارک بوده و هـدف آن‌هـا تشـخیص طبقه موضوعی نیست.

Kanoulas و همکاران با این فرض که توزیـع آمـاری وزن اصطلاحات در مدارک مربوط و نامربوط بـر پایـه یـک مـدل استوار است، پژوهشی را انجام دادند. آن‌ها دریافتند که توزیـع اصطلاحات در مدارک مربوط و نـامربوط بـه صـورت نمـایی است (۱۰).

Hoashi و همکـاران بـا ایـن فـرض کـه در یـک سیسـتم گزینشی اطلاعات می‌توان از مـدارک نـامربوط بهـره بـرد تـا بتوان مدارک مربوط به نیاز کاربر را فیلتر و ارسال کرد، مـدلی را عرضه کردند. در مدل پیشنهادی، یک فهرست ایجـاد شـد که حاوی اطلاعات مدارک نامربوط بود کـه در مرحلـه فیلتـر کردن بازیابی شده بودند. آن‌ها بـر ایـن بـاور بودنـد کـه ایـن فهرست می‌تواند در مراحل بعـد، بازیـابی مـدارک نـامربوط را کاهش دهد به گونـه‌ای کـه سـامانه بازیـابی بتوانـد مـدارک مرتبطتر را بازیابی کند که ممکن است در مرحله وزن‌سـنجی، وزنشان در حد آستانه تعیین شده نباشد (۱۱).

در کنار ادبیات بالا، ادبیات دیگری وجود دارد که به تشخیص و تعیین طبقـه یـا شـاخه موضـوعی پرداختـه اسـت. یکـی از قـدیمی‌تـرین آن‌هـا، اثـر Sparck Jones و Needham است. آن‌ها بر اساس تئوری توده‌گرا، برنامه‌ای طراحی کردنـد که شباهت بین یک جفت اصطلاح را بر اساس رخـداد و هـم رخـدادی آن‌هـا در مـدارک مـی‌سـنجید و اصـطلاحاتی کـه بیشترین شباهت را از این نظـر داشـتند در یـک طبقـه جـای می‌داد (۱۲). در دهه اخیر می‌توان به اثر Collier، Nobata و Tsujii اشاره کرد که به تعیین و طبقه‌بندی اصطلاحات در حوزه زیست شناسی ملکولی پرداختند. آن‌ها با یک جستجوی کنترل شـد در MEDLINE، مجموعـه‌ای از چکیـده‌هـا را ذخیره کردنـد و بـر اسـاس نظـر متخصصـان، کـار تعیـین و طبقه‌بندی اصطلاحات را انجام دادند. نکته‌های اصـلی کـه از نظر آن‌ها در تعیین و طبقه‌بندی این اصطلاحات مـورد توجـه آن‌ها قرار گرفت عبارت بودند از ویژگی‌های واژگانی، دستوری و معنایی اصطلاحات (۱۳).

در اثری دیگر Song و همکارانش، مدل خودکـاری را بـرای طبقه‌بندی مدارک عرضه کردند که بـر پایـه هسـتی‌شناسـی استوار است. در این مدل، اصطلاحات و واژگـان مـدارک وب به شکل سلسله مراتبی بیان می‌شوند. طبقه‌بندی مـدارک بـر پایه هستی‌شناسـی، ویژگـی‌هـایی از مـدارک وب را در نظـر می‌گیرد که می‌تواند آن‌ها را به صحیح‌تـرین شـکل عرضـه و پس از تحلیل محتوای آن‌ها، با توجه به طبقات از پیش تعیین شده برای هر مدرک، مـدرک مـورد نظـر را در مناسـب‌تـرین طبقه گروه‌بندی کند (۱۴).

Utsuro و همکارانش روشی را پیشنهاد کـرده‌انـد تـا بـر آن اساس بتـوان حـوزه خـاص اصـطلاحات تخصصـی را در وب تخمـین زد. فـرض آن روش ایـن اسـت کـه در هـر حـوزه، فهرستی از اصطلاحات شناخته شده وجود دارد. بر اساس ایـن فهرست می‌توان در وب جستجو کرد و بر اساس مـدل بـردار فضایی اصطلاحات، مدارک مربوط به آن حـوزه را تشـخیص داد (۱۵). یکی از اخیرترین آثار مربوط بـه Marin-Castro و همکاران است. آن‌ها به طبقـه‌بنـدی خودکـار پایگـاه‌هـای اطلاعاتی در وب پرداخته‌اند (۱۶). کار آن‌ها از جهاتی بـا کـار Utsuro و همکاران شبیه است. آن‌ها برای این کـار از روش

تشکیل واژگان تخصصی مرتبط با حوزه موضوعی توجه کرده‌اند (۱۵). به طور خلاصه، هدف پژوهش، تعیین کارآمدی روش تشخیص خودکار شاخه اصلی اصطلاحات MeSH از طریق محاسبه نسبت فراوانی آن‌ها در دسته مدارک مرتبط و غیر مرتبط است که از طریق PubMed بازیابی شده‌اند.

روش بررسی

روش مطالعه توصیفی و از مطالعات تحلیل استنادی و نوع مطالعه کاربردی بوده است. در تیر ماه ۱۳۹۱ شمسی، از MeSH و پایگاه PubMed به عنوان منابع گردآوری اطلاعات بهره گرفته شد. این منابع، از دسته منابع مشهور و معتبر در حوزه پزشکی و مورد تاکید و استفاده در بسیار از کتابخانه‌ها و مراکز اطلاع‌رسانی دنیا هستند. جایگاه این منابع، روا بودن بهره‌گیری از آن‌ها را به عنوان ابزار تائید می‌کند. یکی از کارهایی که برای آزمودن کارآمدی روش پیشنهادی می‌توان انجام داد، بهره گیری از طرح طبقه‌بندی سرعنوان‌های موضوعی پزشکی MeSH و دسته‌بندی مدارک PubMed بر اساس سرشاخه‌های اصلی آن است. به گونه‌ای که هر مدرکی که در PubMed ، حداقل یکی از اصطلاحات موضوعی شاخه مورد نظر را در بر دارد، به عنوان مدرک مرتبط به آن شاخه محسوب گردد. سر انجام، تعداد "دسته مدارک" ما به تعداد سر شاخه های درخت سرعنوان های موضوعی MeSH (۱۹ دسته) خواهد رسید. مسلما، هر مدرک ممکن است موضوعاتی از شاخه‌ها یا طبقات مختلف MeSH دریافت کرده باشد که باعث می‌شود هر مدرک بتواند در بیش از یک دسته هم قرار بگیرد. همان طور که آمد، فرض بر این است، دسته‌ای که هر یک از مدارکش حداقل یک اصطلاح مرتبط با شاخه مورد نظر ما را در بر دارد، مرتبط با آن شاخه است. فرض نهایی این است که نسبت فراوانی یک اصطلاح از یک شاخه باید در دسته مدارک مرتبط به آن شاخه، بیشتر از نسبت فراوانی آن در سایر شاخه‌ها باشد.

تعداد اصطلاحات MeSH به بیش از ۲۴۰۰۰ مورد بالغ می شود. تعداد مدارک PubMed که از طریق PubMed قابل جستجو هستند به هجده میلیون می‌رسد. این تعداد اصطلاح موضوعی MeSH و مدارک نمایه سازی شده در PubMed جامعه این پژوهش را تشکیل می‌دهند.

برای استخراج داده‌های لازم از PubMed و MeSH، مراحل عملیاتی زیرصورت گرفت:

۱. با فرمول‌هایی شبیه زیر در PubMed جستجو شد که در هر جستجو قسمت اول، یعنی نام طبقه عوض می‌شد. جمعا یازده جستجو به تعداد طبقات مورد نظر انجام شد:

Anatomy Category"[Mesh] AND 2[Volume] AND Journal Article [ptyp] AND English [lang] AND Medline [sb]

همان‌طور که از فرمول جستجوی بالا بر می‌آید، جستجو در جلد دوم، مقالات نشریات، زبان انگلیسی و وجودشان در MEDLINE محدود شد. دلیل اعمال این محدودیت‌ها این بود که تعداد مدارک مورد بررسی به گونه ای کاهش یابد که در آن سوگیری ایجاد نشود. در صورتی که این محدودیت‌ها اعمال نمی‌شد، تعداد رکوردهای بازیابی شده به کل مدارک موجود PubMed (۱۸ میلیون) می‌رسید که ذخیره و پردازش این حجم از رکورد به دلیل محدودیت‌ها فنی میسر نیست. روایی این عبارات جستجو، از طریق متخصصان علم اطلاعات و دانش‌شناسی مورد تائید قرار گرفت. در اصل، سرعنوان‌های موضوعی MeSH دارای نوزده طبقه است. اما به خاطر مسائلی مانند نپرداختن به محتوا و پرداختن به سرعنوان‌های شکلی، هشت شاخه یا طبقه زیر از پژوهش حذف شدند:

۱) Phenomena and Processes Category
۲) Disciplines and Occupations Category
۳) Persons Category
۴) Pharmacological Actions Category
۵) Publication Type Category
۶) Check Tags Category
۷) Subheadings Category
۸) Geographical Locations Category

۱. رکوردهای بازیابی شده با فرمت MEDLINE روی دیسک سخت ذخیره شد که تعداد فایل‌ها به ۱۱ مورد برابر با طبقات مورد بررسی بالغ می‌شد. تعداد این رکوردها ۴۵۷۹۱۳ مورد بود. اگر تعداد تکرار رکوردها در گروه‌های مختلف در نظر گرفته نشود، در کل، تعداد ۱۶۳۲۲۶ رکورد متمایز ذخیره

شد. این تعداد رکورد، به طور قابل توجه‌ای از میزان حجم نمونه توسط فرمول کوکران بشتر و نزدیک به حجم جامعه بود، لذا همه مدارک بازیابی شده به عنوان نمونه، مورد مطالعه قرار گرفت.

۲. با زبان برنامه نویسی C Sharp برنامه‌ای نوشته شد که اصطلاحات MeSH را از میان رکوردها استخراج می‌کرد. پس از این عمل و کنار گذاشتن چک تاگ‌ها، مشخص گردید که واژگان اصطلاحی رکوردها به ۱۸۱۶۴ مورد می‌رسد که در مقایسه با کل سرعنوان‌های MeSH یعنی ۲۴،۰۰۰، مقدار قابل قبولی است و از محاسبه حجم نمونه توسط فرمول کوکران به طور قابل ملاحظه ای بالاتر است..

۲. با زبان برنامه نویسی C Sharp، برنامه دیگری نوشته شد که نقش ربات داشت. یعنی هر یک از این ۱۸۱۶۴ اصطلاح را یکی به یکی در نسخه اینترنتی سرعنوان‌های موضوعی MeSH به صورت خودکارجستجو می‌کرد و با بررسی و تجزیه خودکار صفحه HTML بازیابی شده، طبقه اصلی آن اصطلاح را مشخص می‌کرد.

۳. برنامه دیگری با زبانC Sharp نوشته شد که نسبت رخداد هر اصطلاح را در هر طبقه از مدارک با فرمول زیر مشخص می کرد:

$$tof_{ij} = \sum_{ij} t / \sum_{i} t$$

در فرمول بالا، tof برابربا بسامد نسبی هر اصطلاح در مدارک هر یک از طبقات یازده گانه، t نشانگر اصطلاح MeSH، i نشانگر چندمین از ۱۸۱۶۴ اصطلاح و j نشانگر چندمین طبقه از یازده طبقه است.

۴. پس از تعیین بسامد نسبی هر اصطلاح در هر طبقه، مشخص شد که بسامد آن اصطلاح در مدارک کدام طبقه بیشتر بوده است.

تمامی این مراحل شش گانه به این منظور انجام شد که داده‌های لازم برای پاسخ به پرسش‌های یک تا چهار این پژوهش فراهم شود. مرحله دیگری که صورت گرفت برای تهیه داده‌های مربوط به پرسش پنجم و ششم بود که فهرست اصطلاحات MeSH در قالب یک جدول برای متخصصان پزشکی ارسال شد و از آن‌ها خواسته شد که تعیین کنند، هر اصطلاح به کدام یک از یازده طبقه موضوعی MeSH تعلق دارد. تعداد اصطلاحات ارسالی برای نظرخواهی از متخصصان، یک درصد از کل اصطلاحات بود که به صورت تصادفی انتخاب شدند. دلیل انتخاب یک درصد این بود که در عمل نظر خواهی درباره ۱۸۱۶۵ اصطلاح توسط انسان ممکن نیست. به عبارتی، متخصصان حاضر نمی‌شوند که کار خود را رها کنند و مدت زمان بسیار زیادی را (که بالغ بر هفته‌ها خواهد شد) به تشخیص طبقه موضوعی اصطلاحات برای پژوهش دیگران بپردازند.

پس از تعیین نسبت فراوانی هر یک از اصطلاحات مورد بررسی در یازده دسته از مدارک بازیابی شده و تعیین صحت پیش‌بینی برای تشخیص خودکار شاخه هر یک از این اصطلاحات بر اساس نسبت فراوانی در دسته مدارک مرتبط و نامرتبط، از برنامه‌هایExcel و SPSS بهره گرفته شد. برای یافتن تعداد طبقاتی که هر اصطلاح به آن تعلق دارد و همین‌طور برای تعیین این که هر مدرک در چند طبقه جای می‌گیرد، توزیع فراوانی آن‌ها محاسبه گردید.

از آزمون تحلیلیT برای تعیین معنادار بودن تفاوت نتایج حاصل از روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط و غیر مرتبط با نتایجی که به طور تصادفی ممکن بودن حاصل شود و هم‌چنین تفاوت نتایج حاصل از طبقه‌بندی متخصصان پزشکی با روش یاد شده و طبقه پیشنهادی هر اصلاح توسط MeSH استفاده شد. برای روشن شدن این که روش پیشنهادی تحت تاثیر موضوعی اصلی مدارک است، از آزمون Chi-Squar بهره گرفته شد.

یافته‌ها

نتایج آماری این پژوهش در دو بخش قابل ارائه است:

۱. یافته‌های توصیفی که به صورت آماری، به پیش فرض‌های این پژوهش (پرسش ۱ و ۲) توجه دارد.

۲. یافته های تحلیلی که به پاسخ به پرسش های ۴ تا ۶ می پردازند.

یافته‌ها در رابطه با فراوانی حضور هر مدرک در طبقات مختلف نشان داد که مدارک معمولا در بیش از یک طبقه حضور دارند.

نمودار ۱ نشان می‌دهد که هر مدرک بازیابی شده از PubMed، احتمالا به ۱ تا ۱۰ طبقه تعلق دارد. بر اساس نمای به دست آمده، به طور متوسط هر مدرک می‌تواند به سه طبقه تعلق داشته باشد. با توجه به داده‌های موجود، اگر تکرار مدارک مورد بررسی در نظر گرفته نشود، در این پژوهش، تعداد ۱۶۳۲۲۶ مدرک از PubMed بازیابی شده است. اگر این تکرار در نظر گرفته شود، تعداد مدارک بازیابی شده از یازده جستجوی انجام شده، ۴۵۷۹۱۳ مورد است. با محاسبه نسبت این دو مقدار، میانگین طبقات موضوعی مدارک بازیابی شده از PubMed **به** دست خواهد آمد. این نسبت در پژوهش حاضر برابر با ۲/۸۱ است که با نمای به دست آمده در نمودار بالا، نزدیکی فراوانی دارد.

$$\text{میانگین تعداد طبقات یک مدرک} = \frac{\text{مدارک بازیابی شده با احتساب تکرار}}{\text{مدارک بازیابی شده بدون احتساب تکرار}}$$

نمودار ۱: فراوانی مدارک بر اساس تحت پوشش قرار دادن طبقات موضوعی Mesh

در ادامه، یافته‌های مربوط به حضور هر اصلاح در دسته‌های مختلف مدارک نشان داد که تقریبا امکان حضور هر اصطلاح MeSH در تمامی یازده دسته مدرک وجود دارد.

نمودار ۲ نشان می‌دهد، از ۱۸۱۶۴ اصطلاح MeSH مورد مطالعه، تعداد ۱۱۳۸۹ مورد آن در هر یازده دسته از مدارک بازیابی شده مشاهده شده است. یعنی اکثریت اصطلاحات موضوعی در تمامی شاخه‌های زیست پزشکی به کار رفته است.

یافته‌های حاصل از آزمون تحلیلیT نشان داد که روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط و غیر مرتبط، احتمال تشخیص خودکار شاخه اصلی اصطلاحات MeSH را به طور معنادارافزایش می‌دهد. جدول ۱، یک ماتریس مربع شکل است که ستون‌ها و سطرهای آن متناظر هستند؛ مثلا عنوان ستون اول و سطر اول از این جدول نظیر هم و برابر با Analytical است و بقیه ستون‌ها و سطرها نیز به این گونه نظیرند. اعداد موجود در قطر این جدول با رنگ خاکستری مشخص شده و نشانگر تعداد اصطلاحاتی است که روش پیشنهادی توانسته است منطبق برساختار درختی سرعنوان های موضوعیتMeSH تشخیص بدهد. به عبارتی، نشانگر فراوانی مشترک روش پیشنهادی پژوهش و دست اندرکاران ایجاد و توسعه سرعنوان‌های موضوعی MeSH برای تشخیص طبقات موضوعی اصطلاحات است. جمع

ستون‌های این جدول، نشان می‌دهد که در کل چه تعداد از اصطلاحات توسط دست اندرکاران ایجاد و توسعه سرعنوان‌های موضوعی MeSH به عنوان زیرطبقه‌های یک طبقه اصلی تشخیص داده شده است. در برابر، جمع هر یک از سطرها، نشان می‌دهد که چه تعداد از اصطلاحات توسط روش پیشنهادی، به عنوان زیر طبقه‌های یک طبقه اصلی تشخیص داده شده است. مثلا در جمع ستون اول، عدد ۲۱۹۹ حاصل شده است؛ یعنی دست اندرکاران سرعنوان‌های موضوعی MeSH تشخیص داده‌اند که طبقه اصلی ۲۱۹۹ اصطلاح مش، طبقه Analytical است. اگر جمع اولین سطر را در نظر گرفته شود، عدد ۲۸۹ حاصل شده است؛ یعنی روش پیشنهادی، طبقه اصلی این تعداد از اصطلاحات را Analytical تشخیص داده است. ستون اول از همین سطر بیانگر این است که ۹۲ اصطلاح موضوعی در حوزه Analytical است که بین روش پیشنهادی و دست‌اندرکاران سرعنوان موضوعی MeSH مشترک تشخیص داده شده‌اند.

می‌توان گفت جدول ۱ در یک نگاه کلی ۱۱طبقه موضوعی که در این پژوهش بدان‌ها پرداخته شده را به تفکیک مورد مقایسه قرار می‌دهد و فراوانی هر طبقه را در سه مورد بررسی می‌کند. در مورد اول فراوانی هر طبقه را در سرعنوان موضوعی MeSH، در مورد دوم فراوانی هر ۱۱ طبقه را در روش پیشنهادی و مورد سوم موارد مشترک را در مدل پیشنهادی و سرعنوان‌های موضوعی MeSH ارائه می‌دهد که همان قطر جدول مذکور می‌باشند.

انتظار نمی‌رود که تشخیص شاخه‌ی موضوعی یک اصطلاح صد در صد با ساختار MeSH منطبق باشد اما احتمال آن را می‌توان پیش بینی کرد. چون در این پژوهش یازده دسته مدرک از جستجوی سرشاخه موضوعات حاصل شده است و به عبارتی، چون یازده طبقه موضوعی داریم، از لحاظ قاعده احتمالات این که شاخه اصلی یک اصطلاح MeSH به طور تصادفی درست تشخیص داده شود برابر با یک یازدهم ($\frac{1}{11}$) است.

برای دریافتن این که آیا تعداد اصطلاح مشترک مشاهده شده تصادفی بوده یا ناشی از توان روش محاسبه بسامد رخدادی است، یک آزمون T جفتی انجام شد. در این آزمون، تفاوت میانگین ستون چهارم یعنی تعداد اصطلاحات مشترک تعداد اصطلاحات مشترک در حالت تصادفی (پیش بینی) با ستون پنجم جدول یعنی، تعداد تعداد اصطلاحات مشترک (مشاهده شده)اندازه‌گیری شد.

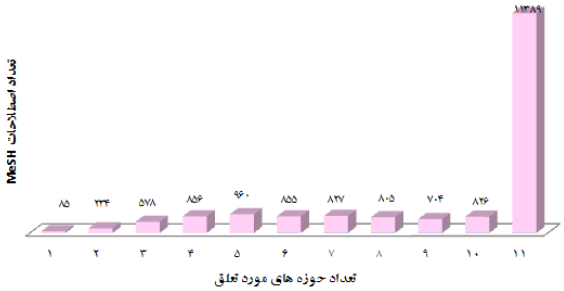
نتیجه این مقایسه نشان داد که مقدار T برابر با ۲/۳۶۹- و درجه آزادی ۱۰ و معنا داری ۰/۰۳۹ است. چون معناداری از ۰/۰۵ کمتر است، با اطمینان ۹۵ درصد می‌توان استنباط کردکه روش محاسبه بسامد رخدادی اصطلاحات در مدارک مربوط و نامربوط، احتمال تشخیص صحیح طبقه موضوعی یک اصطلاح MeSH را به طور معنادار افزایش می‌دهد.

آزمون Chi-Squar نشان داد که عملکرد روش مقایسه نسبت فراوانی اصطلاحات در مدارک مرتبط وغیر مرتبط برای تشخیص خودکار شاخه اصلی اصطلاحات مش، در همه شاخه‌های MeSH به یک اندازه نیست.

نگاه دقیق‌تر به یافته‌ها در جدول ۲ نشان می‌دهد که درصد تاثیر روش پیشنهادی این پژوهش در طبقات مختلف موضوعی از ۳ درصد تا ۶۷ درصد متفاوت است. این یافته به این معنا است که احتمالا، بین طبقه موضوعی به عنوان یک متغیر و میزان تاثیر روش پیشنهادی، رابطه‌ای وجود دارد. به عبارتی، توان این روش در تشخیص طبقه موضوعی یک اصطلاح، به شاخه موضوعی آن بستگی دارد. اگر چنین باشد، در این روش می‌توان شاخه موضوعی را به عنوان یک متغیر مداخله‌گر محسوب کرد. برای بررسی نقش این متغیر، از آزمون Chi-Squar بهره گرفته شد:

نتیجه آزمون Chi-Squar به میزان ۲۰۵۷/۲ و درجه آزادی ۱۰ نشان داد که معناداری رابطه بین طبقه موضوعی به عنوان یک متغیر مداخله‌گر و میزان تاثیر روش پیشنهادی، بالای ۹۹ درصد است. بنابراین، با اطمینان بالا می‌توان گفت که میزان تاثیر گذاری روش پیشنهادی تابع طبقه موضوعی مورد بررسی است. یعنی در طبقات مختلف میزان تاثیر گذاری

متفاوت است. بر اساس نتیجه آزمون تحلیلی T، تخمین متخصصان پزشکی درباره طبقه موضوعی اصطلاحات، به طور معناداری بر ساختار سرعنوان‌های موضوعی MeSH منطبق است. یکی از راه‌هایی که در این پژوهش برای سنجش مدل پیشنهادی در نظر گرفته شد، این بود که تشخیص متخصصان علم پزشکی را در زمینه طبقه اصلی اصطلاحات جویا شویم. در این مورد، مقدار T جفتی برابر با ۲۵/۸۱۰ و تعداد نمونه‌ها برابر با ۱۵۲ و درجه آزادی ۱۵۱ و سطح معناداری از ۰/۰۰۰ کمتر بود، بنابراین، تشخیص هر سه متخصص به صورت معناداری بر ساختار طبقات سرعنوان‌هایMeSH منطبق است. بار دیگر از آزمون T جفتی استفاده شد تا این بار انطباق تشخیص متخصصان پزشکی، با تشخیص روش پیشنهادی سنجیده شود. بعد از جمع آوری نظر سه متخصص موضوعی، نتیجه تشخیص آن‌ها با انجام آزمون T جفتی سنجیده شد و در این مورد، تعداد نمونه‌ها برابر با ۱۵۲ و درجه آزادی ۱۵۱ و سطح معناداری ۰/۴۹ بود. بنابراین فرض برابری یا همخوانی بین نظر متخصصان و مدل پیشنهادی این پژوهش تایید نمی‌شود.



نمودار ۲: فراوانی تعلق اصطلاحات Mesh به حوزه‌های مختلف موضوعی

جدول ۱: فراوانی اصطلاحات در شاخه‌های سرعنوان‌های موضوعیMesh و روش پیشنهادی

	فراوانی در سرعنوان‌های موضوعیMESH												
شاخه موضوعی	Analytical	Anatomy	Chemicals	Diseases	Health Care	Humanities	Organisms	Psychiatry	Information	Anthropology	Technology & Food	جمع (روش پیشنهادی)	فراوانی در روش پیشنهادی
Analytical	۹۲	۱	۱۱۳	۵۹	۳	۰	۱۹	۲	۰	۰	۰	۲۸۹	
Anatomy	۱۷۲	۵۸۹	۷۵۴	۳۴۵	۱۰	۱	۱۳۶	۹	۶	۵	۴	۲۰۳۱	
Chemicals	۲۴۷	۲۱۳	۱۵۸۱	۲۹۹	۳۳	۴	۲۸۷	۲۰	۱۶	۱۱	۱۵	۲۷۲۶	
Diseases	۱۹۰	۴۷	۴۴۱	۷۳۶	۸	۱	۱۲۰	۷	۵	۳	۴	۱۵۶۲	
Health Care	۵۳۵	۱۵۵	۹۶۷	۷۱۴	۱۷۰	۴	۳۴۲	۷۸	۳۱	۴۶	۱۹	۳۰۶۱	
Humanities	۲۸۷	۱۵۰	۱۰۲۷	۵۱۶	۱۰۵	۱۰۴	۲۸۹	۱۱۶	۴۴	۹۳	۱۵	۲۷۴۶	
Organisms	۰	۱	۸۰	۴۹	۵	۰	۴۵	۰	۰	۰	۰	۱۸۰	
Psychiatry	۱۰۸	۸۱	۴۴۵	۲۹۱	۳۸	۲	۹۱	۳۲۱	۱	۱	۰	۱۳۷۹	
Information	۱۵۱	۳۹	۴۸۴	۱۳۵	۴۷	۲	۱۲۰	۲۶	۸۵	۵	۶	۱۱۰۰	
Anthropology	۱۰۹	۳۱	۲۲۵	۱۷۳	۱۸۷	۱	۸۴	۵۱	۲	۱۷۱	۳	۱۰۳۷	
Technology & Food	۳۰۸	۹۱	۹۰۸	۳۳۳	۶۸	۶	۱۷۲	۳۰	۷	۱۱	۱۱۹	۲۰۵۳	
جمع(درMeSH)	۲۱۹۹	۱۳۹۸	۷۰۲۵	۳۶۵۰	۶۷۴	۱۲۵	۱۷۰۵	۶۶۰	۱۹۷	۳۴۶	۱۸۵	۱۸۱۶۴	

جدول ۲: فراوانی انطباق شاخه موضوعی اصطلاحات با سرعنوان موضوعی مش

طبقه	تعداد اصطلاحات مشاهده شده در MeSH	احتمال تصادفی بودن تعداد اصطلاحات مشترک	تعداد اصطلاحات مشترک در حالت تصادفی(پیش بینی)	تعداد اصطلاحات مشترک در حالت واقعی (مشاهده شده)	درصد اصطلاحات مشترک (مشاهده شده)
Analytical	۲۱۹۹	$\frac{۱}{۱۱}$	۱۹۹/۹	۹۲	۴٪
Anatomy	۱۳۹۸	$\frac{۱}{۱۱}$	۱۲۷/۰۹	۵۸۹	۴۲٪
Chemical& Drugs	۷۰۲۶	$\frac{۱}{۱۱}$	۶۳۸/۷۲	۱۵۸۱	۲۳٪
Diseases	۳۶۵۰	$\frac{۱}{۱۱}$	۳۳۱/۸۱	۷۳۶	۲۰٪
Health Care	۶۷۴	$\frac{۱}{۱۱}$	۶۱/۲۷	۱۷۰	۲۵٪
Humanities	۱۲۵	$\frac{۱}{۱۱}$	۱۱/۳۶	۱۰۴	۶۷٪
Organisms	۱۷۰۵	$\frac{۱}{۱۱}$	۱۵۵	۴۵	۳٪
Psychiatry	۶۶۰	$\frac{۱}{۱۱}$	۶۰	۳۲۱	۴۹٪
Technology& Food	۱۸۵	$\frac{۱}{۱۱}$	۱۶	۱۱۹	۶۴٪
Anthropology	۳۴۶	$\frac{۱}{۱۱}$	۳۱/۴۵	۱۷۱	۴۹٪
Information Science	۱۹۷	$\frac{۱}{۱۱}$	۱۷/۹	۸۵	۵۱٪

بحث

به اعتقاد Hjørland و Albrechtsen «سامانه‌های اطلاعاتی تحت تاثیر حوزه‌های دیگر مانند علوم شناختی، زبانشناسی، روانشناسی، مطالعات آموزشی، علوم رایانه، جامعه‌شناسی و فلسفه هستند» (۱۷) که برخی از این حوزه‌ها ذاتا عینی و برخی ذاتا ذهنی هستند. از طرفی ساختار درختی MeSH بر اساس یک مدل نظری و انسان‌مدار (ذهنی) استوار است در حالی که مدل پیشنهادی یک مدل ماشین مدار یا کمیت محور (عینی) است. از این رو، شایسته است که به بحث حاضر از زاویه عینی بودن و ذهنی بودن پرداخته شود.

شاید از جنبه نظری، یک اصطلاح موضوعی در طبقه‌ای خاص جای داده شود اما به این معنا نیست که تنها مورد استفاده متخصصان شاخه تعیین شده است. نگاه به پاسخ پرسش‌های اول و دوم این گفته را تایید می‌کند زیرا همان گونه که مشاهده شد، هر مدرک در علوم پزشکی به طور متوسط به سه شاخه موضوعی تعلق دارد و هر اصطلاح موضوعی در این حوزه می‌تواند در مدارک کلیه شاخه‌های موضوعی علوم پزشکی ظهور یابد. یافته مرتبط با این دو پرسش، منطبق با گفته Mai است که به ساده نبودن «مرزبندی رشته‌ها»اعتقاد دارد (۱). پاسخ به پرسش دوم نشان داد که اکثر اصطلاحات موضوعیMeSH توسط متخصصان همه شاخه‌های علوم پزشکی مورد استفاده قرار می‌گیرند. مسلما در رشته‌هایی مانند ادبیات فارسی و مهندسی معدن که افتراق موضوعی زیادی دارند، میزان اصطلاحات موضوعی آن‌ها در این سطح نیست اما حدس زده می‌شود که قابل ملاحظه باشد. از این رو، برای روشن‌تر شدن لایه‌های پنهان این مساله، آزمودن شباهت‌های موضوعی رشته‌هایی که از جنبه نظری کاملا از هم مجزا هستند، لازم باشد. در پاسخ به پرسش‌های سوم و چهارم مشاهده شد که روش پیشنهادی به عنوان یک روش ماشینی و عینی، احتمال تشخیص منطبق بر تشخیص ماشین را افزایش می‌دهد اما در برخی از شاخه‌ها میزان انطباق اندک و حداکثر انطباق یا هم خوانی ۶۷ درصد است. این نشان می‌دهد که در برخی از حوزهای دانش، سطح انطباق ساختار و تقسیم‌بندی موضوعات در دو روش عینی و ذهنی بالاتر و در برخی پایین‌تر است. به گفته Kwasnik «با روش‌های متفاوتی طرح‌های طبقه بندی و دانش در تعامل با یکدیگر قرار می‌گیرند. گاهی این تعامل آن قدر موزون و هماهنگ است که این دو مدت زمان زیادی در پیوند با هم هستند.

گاهی دانش تغییر می‌کند و طبقه‌بندی نیز باید تغییر کند یا دانش تغییر می‌کند و طبقه‌بندی دیگر با آن منطبق نیست. گاهی طبقه‌بندی خودش باعث تولید دانش جدید می‌شود»(۱۸). توکلی‌زاده‌راوری و نجابتیان که هم‌نشینی اصطلاحات MeSH در زمینه روانشناسی ازدواج را با روش خوشه بندی سلسله مراتبی، در دو دهه مختلف مقایسه کرده بودند، دریافتند که برخی اصطلاحات هستند که دائما در یک دسته و خوشه قرار می‌گیرند در حالی که برخی از اصطلاحات خوشه موضوعی ثابتی ندارند. به عنوان مثال، موضوعات افسردگی و اضطراب یا موضوعات جنسی ازدواج، اصطلاحاتی بودند که در دوره بیست ساله مورد مطالعه همیشه با هم و در یک خوشه بودند و موضوعی مثل شاخت درمانی هم نشینان یا خوشه ثابتی نداشت (۱۹).

با توجه به مباحث بالا و یافته پرسش چهارم این پژوهش، می‌توان گفت که روش‌های نیمه هدایت شده که اساس آن‌ها تعیین شاخه‌ها و طبقات اصلی از قبل است و ماشین کار هدایت مدارک یا اصطلاحات را به طبقه‌ای خاص انجام می‌دهد، برای همه شاخه‌ها نمی‌تواند مناسب باشد. به عبارتی، در برخی از شاخه‌های موضوعی مفاهیم آن‌ها در نظر و عمل به هم نزدیک و در تعامل با یکدیگر هستند و در برخی از شاخه‌ها از هم دورترند. احتمالا، این دوری در شاخه‌های موضوعی بین رشته‌ای بیشتر است. از این رو، روش نیمه هدایت شده برای این گونه شاخه‌های موضوعی نمی‌تواند گویای واقعیت باشد و باید از روش‌های هدایت نشده برای طبقه بندی آن‌ها بهره برد چون به گفته Albrechtsen و Jacob باعث «عدم هم‌خوانی بین ساختار طرح‌های طبقه‌بندی در سامانه‌های بازیابی اطلاعات و ساختار دانش افراد و راهبردهای جستجو آن‌ها» می‌شود (۲۰).

در پاسخ به پرسش‌های پنجم و ششم دیده شد که نظر متخصصان علوم پزشکی در خصوص طبقه موضوعی اصطلاحات بر مدل MeSH منطبق بود و با مدل پیشنهادی انطباق کمتری داشت. در حالی که مدل پیشنهادی در این پژوهش، واقعیت عملی صورت گرفته را بیان می‌کند و نشان می‌دهد که ساختار ذهنی موضوعات تا چه اندازه از عمل دور است. حال چه می‌شود که بین نظر و عمل تفاوت ایجاد می‌شود. نمونه زیر این مساله را با یک توضیح عملی روشن می‌سازد:

در پژوهش حاضر، متخصصان علوم پزشکی اصطلاح کلسیم را یک ماده شیمیایی با کاربرد دارویی تشخیص دادند، در صورتی که اگر بخواهند یک مقاله در مورد دیواره‌های سلولی بنویسند، به اصطلاح کلسیم به عنوان یک ماده تشکیل دهنده غشای سلولی نگاه خواهند کرد. نمونه دیگر، اصطلاح آهن است که متخصصان به عنوان موضوعی در زمینه شیمی و دارو تشخیص دادند، در حالی که در مدل پیشنهادی ما و بعد از سنجش نسبت بسامد رخداد، در حوزه تکنولوژی و غذا قرار گرفت. مسلما، اصطلاح آهن را می‌توان به عنوان یک ماده اصلی مورد نیاز بدن که در غذاها یافت می‌شود، نیز در نظر گرفت.

Sinclair نمونه‌ای مشابه بیان می‌دارد و سپس نشان می‌دهد که چگونه یک موضوع می‌تواند غالبا در چندین طبقه گنجانیده شود. او در چرایی این اتفاق، به طبقه‌بندی به عنوان یک امر مربوط به ادراک و شناخت نگاه می‌کند تا به قول خودش بفهمد «چرا طبقه‌بندی چیزها در این حالت، غالبا دشوار است». به این منظور، او معتقد است که باید بین طبقه بندی و رده بندی تفاوت قائل شد (۲۱). وی از قول Jacob (۲۲) بیان می‌دارد که طبقه‌بندی، «فرآیند تقسیم جهان به گروه‌هایی از اشیا است که اعضای آن به گونه‌ای به هم شباهت دارند» و «طرح‌های رده بندی، دسته‌ای از رده‌های منحصر و بدون هم پوشانی هستند که در یک ساختار سلسله مراتبی تنظیم شده‌اند و منعکس کننده یک نظم واقعی از پیش تعیین شده هستند». در ادامه با بیان تصویری که Weinberger (۲۳) ارائه کرده است، رده‌بندی را یک درخت می‌داند که هر برگش به یک شاخه و هر شاخه‌اش به شاخه دیگر و نهایتا شاخه‌ها به تنه چسبیده‌اند. در برابر، طرح‌های طبقه‌بندی توده‌ای از برگ‌ها هستند. از این جهت، طبقه‌بندی مفهومی اعم بر رده‌بندی است. با تکیه بر این

نظریات می‌توان نتیجه گرفت، آن چه که ماشین می‌تواند انجام بدهد، تشخیص طبقه یا طبقات یک اصطلاح موضوعی است که بر اساس فنون مختلف ریاضی، آماری، دستور زبانی و غیره انجام می‌دهد. از آن جا که رده‌بندی کاری کاملا ذهنی و ادراکی و به قول Kwasnik «خوشه‌بندی تجربه‌های بشری است»(۱۸)، مدل‌های ماشینی نمی‌توانند دقیقا آن فرآیند را مشابه سازی کند و به این جهت در بازیابی اطلاعات هم نمی‌توانند جستجوی کاربر را با آن تطبیق دهند.

نتیجه‌گیری

این پژوهش بر پایه این فرض اساسی استوار است که هر اصطلاح موضوعی به طبقه یا شاخه‌ای تعلق دارد که نسبت حضورش در آن بالاتر است. این فرض ناشی از آن است که اصطلاحات موضوعی پویا هستند و این برخلاف روش‌های رده بندی است که هر اصطلاح را تنها متعلق به یک رده می‌دانند.

یافته‌های این پژوهش نشان داد که علاوه بر اصطلاحات موضوعی، مدارک هم پویا هستند و هر مدرک می‌تواند به طور متوسط به سه شاخه از سرعنوان‌های موضوعی MeSH تعلق داشته باشد و یافته دیگر شاهدی بر این مدعا است که موضوعات را نمی‌توان به شاخه یا طبقه خاصی محدود کرد، زیرا اکثر اصطلاحات مورد مطالعه در همه شاخه‌های موضوعی به کار رفته بودند.

یافته‌های دیگر نشان می‌دهد که انطباق روش‌های ذهنی و عینی همه جا یکسان نیست و این مساله به طور معناداری به شاخه موضوعی مربوط است. از این جهت، روش‌های نیمه هدایت شده برای دسته بندی اصطلاحات و مدارک، در برخی از شاخه‌های موضوعی مناسب است و در شاخه‌هایی مناسب نیست و هم‌چنین بین نگاه ذهنی و عینی برای تشخیص شاخه مرتبط به موضوع، تفاوت معناداری وجود دارد.

References

1. Mai JE. Semiotics and Indexing: an Analysis of the Subject Indexing Process. Journal of Documentation 2005; 57: 567-622.
2. Krauthammer M, Nenadic G. Term Identification in the Biomedical Literature. J. Biomed. Inform 2004; 7(6): 512–526.
3. Vu T and Aw AT, Zhang M. Term Extraction through Unithood and Termhood Unification. Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08); 2008, Hyderabad, India.
4. Harter SP. A probabilistic Approach to Automatic Keyword Indexing. Part II: An algo-rithm for probabilistic indexing. JASIS 1975; 26: 280–9.
5. Kageura K, Umino B. Methods of Automatic Term Recognition. Terminology 1996; 3(2):259.
6. Bookstein A, Swanson DR. Probabilistic Methods for Automatic Indexing. Journal of the American Society for Information Science 1974; 25(5): 312 –18.
7. Bookstein A, Swanson DR. A decision Theoretic Foundation for Indexing. Journal of the American Society for Information Science 1975; 26(1): 45 – 50.
8. Cooper WS, Maron ME. Foundation of Probabilistic and Utility-Theoretic Indexing. Journal of the Association for Computing Machinery 1978; 25: 67-80.
9. Salton G. Automatic Text Processing, Reading. UK: Addison-Wesley; 1989.
10. Kanoulas E, Pavlu V, Dai K, Aslam JA. Modeling the Score Distributions of Relevant and Non-relevant Documents. In ICTIR, Lecture Notes in Computer Science 2009; 5766: 152–63.
11. Hoashi K, Matsumoto K, Inoue N, Hashimoto K. Document Filtering Method using Non-relevant Information Profile. Proceedings of the 23rdAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2000 July 24-28, Athens, Greece: 176-83.
12. Sparck JK, Needham RM. Automatic Term Classification and Retrieval. Information Processing & Management 1968; 4(1): 91–100.
13. Collier N, Nobata C, Tsujii J. Automatic Acquisition and Classification of Terminology using a Tagged Corpus in the Molecular Biology Domain. Terminology 2002; 7(2): 239-57.
14. Song MH, Lim SY, Kang DJ, Lee SJ. Automatic Classification of Web Pages Based on the Concept of Domain Ontology. Proceeding of the 12th Asia-Pacific Software Engineering Conference; 2005.

15. Utsuro T, Kida M, Tonoike M, Sato S. Towards Automatic Domain Classification of Technical Terms: Estimating Domain Specificity of a Term using the Web. In Ng HT; Leong MK, Kan MY, Ji DH, eds. Information Retrieval Technology. New York: Springer; 2006. pp. 633-41
16. Marin-Castro HM, Sosa-Sosa VJ, Lopez-Arevalo I, Escalante-Baldera HJ. Automatic Classification of Web Databases using Domain-Dictionaries. In: Machine Learning and Data Mining in Pattern Recognition. New York: Springer; 2013. pp. 340-51.
17. Hjørland B, Albrechtsen H. Toward a New Horizon in Information Science: Domain-Analysis. JASIS 1995; 46(6): 400-25.
18. Kwasnik, BH. The Role of Classification in Knowledge Represantation and Discovery. Library trends 2000; 48(1): 22-47.
19. Tavakolizadeh-Ravari M, Nejabatian M. Document-Term Clustering: Proximity of Subjects Correspond with Psychology of Marriage in Biomedicine Literature during the Years “1990-99” and “2000-2008. Health Information Management 2010; 7(1-2):172-86. [In Persian]
20. Albrechtsen H, Jacob EK. The Dynamics of Classification Systems as Boundary Objects for Cooperation in the Electronic Library. Library trends 1998; 47(2): 293-312.
21. Sinclair, James. Categorization in Knowledge Contexts, The Australian National University, Department of Engineering. [On Line]. 2006. Available from: URL: http://jrsinclair.com/academic/categorisation-knowledge-contexts
22. Jacob EK. Classification and Categorization: A Difference that Makes a Difference. Library Trends 2004; 52(3):515-40.
23. Weinberger D. Taxonomies and Tags: From Trees to Piles of Leaves. Esther Dyson's Release *10* 2005; 23(2): 1-33.

Automatic Category Recognition of Medical Subject Heading Terms through Comparison of their Occurrence Frequency in Relevant and non-Relevant Documents*

Mohammad Tavakolizadeh-Ravari[1], Saeed Ghaffari[2], Forough Mostafavi[3]

Original Article

Abstract

Introduction: Due to dynamic of terms, their classification is challenging. The current research aims at determining the usability of a model for automatic recognition of MeSH terms categories through measuring their occurrence frequency within relevant and non-relevant document corpuses from PubMed.

Methods: This is a descriptive research that uses the document analysis method. MeSH and PubMed were used to collect research data. The significancy of these resources confirms their validity. 18164 MeSH-term and 163226 PubMed documents were selected. The both of these amounts are greater than what Cocran function suggests. Eleven document corpuses were retrieved from PubMed. The relative ocurrence frequencies of MeSH terms within each corpus were determined. The results were compared with the real category of MeSH. In additions, the categories of 1 percent of MeSH terms were determined by experts in medical domains. The frequency distribution method was used for statistical description of data. Data were also analyzed through T and Chi-Squar tests in SPSS.

Results: Each document of PubMed on average belongs to three MeSH categoris and most of Mesh terms occurred in all corpuses. The results confirm that the suggested method increases the probability of MeSH category recognition. The performance of the method depends on the subject category of MeSH Term and ranges between 3 to 67 percent. The findings also show that the medical expertises determination on the subject category of MeSH Terms is compatible with the real categories of MeSH tree.

Conclusion: The compatibility of the subjective and objective methods for the subject category recognition depends on the knowledge area. The subjective categorization is a quite cognitive task and roots in human environmental experiences. This is why the machine depended models are not able to simulate that process.

Keywords: Classification; Automatic Data Processing; Medical Subject Headings; PubMed.

Received: 4 Mar, 2014 Accepted: 5 Jul, 2014

Citation: Tavakolizadeh-Ravari M, Ghaffari S, Mostafavi F. **Automatic Category Recognition of Medical Subject Heading Terms through Comparison of their Occurrence Frequency in Relevant and non-Relevant Documents**. Health Inf Manage 2015; 12(1):60.

*- This paper is result of an independent research.

1- Assistant Professor, Library and Information Science, Yazd University, Yazd, Iran (Corresponding Author)Email:tavakoli@yazd.ac.ir
2- Assistant Professor, Library and Information science, Payam Noor Qom University, Qom, Iran
3- MSc, Library and Information Science, Payam Noor Eqlid University, Eqlid, Iran