گزارش نهایی طرح پژوهشی

عنوان

تحلیل و تطبیق میزان همخوانی دامنه و اهداف موضوعی نشریات فارسی وزارت عتف با محتوای مقالات منتشر شده در انها طی بازه زمانی سه ساله ۱۳۹۷–۱۳۹۹

Analyzing and matching the scope of MSRT Persian Journals with the content of the articles published in them during the period of 2019-2021

مجری:

دکتر نرجس ورع

تیر ماه ۱۴۰۲


بسم الله الرحمن الرحيم

چکیده

هدف: بر پایهٔ توصیه کمیسیون نشریات علمی در مورد پایبندی به محورهای موضوعی اعلام شده از سوی نشریات علمی، همهٔ نشریات باید به سوی این خط مشی حرکت کرده و انتظار می‌رود محتوای مقالات، انعکاس دهنده پژوهش های مرتبط و مندرج با دامنه و اهداف نشریه باشد. هدف این رویکرد، فراهم ساختن بستر، به گونه‌ای است که جامعهٔ دانشگاهی اعم از پژوهشگران، نویسندگان مقالات و خوانندگان بتوانند هر نشریه را با موضوع یا موضوعات تخصصی شناسایی کنند. از اینرو در پژوهش حاضر تلاش بر این است که با استفاده از روش تحلیل محتوای مقالات نشریات معتبر وزارت علوم در شش حوزه موضوعی سطح کلان، به تعیین میزان همخوانی مقالات با اهداف و دامنه موضوعی اعلام شده از سوی نشریات پرداخته شود.

روش: پژوهش از نوع کاربردی و به روش متن‌کاوی، با رویکرد تحلیلی انجام شده است. جامعه آماری پژوهش حاضر را ۱۱۵۰عنوان نشریه فارسی مصوب وزارت علوم، پژوهش‌ها و فناوری (۱۴۰۰) بر اساس پرتال نشریات[1] تشکیل می‌دهد. باتوجه به محدودیت‌های زمانی، منابع انسانی و مالی، امکان بررسی اهداف و دامنه موضوعی مقالات تمام نشریات در پژوهش حاضر وجود نداشت؛ بنابراین بایستی حجم نمونه مشخص میشد. در همین راستا، حداقل حجم نمونه مورد نیاز توسط نرم افزار، ۱۱۶ عنوان نشریه تعیین شد. در ادامه با استفاده از روش نمونه‌گیری تصادفی طبقه‌ای، نمونه پژوهش به گونه‌ای انتخاب شد که تناسب بین نمونه و جامعه در هر یک از شش حوزه موضوعی سطح کلان و میانی تعیین شده توسط وزارت عتف رعایت گردد. پس از استخراج داده های کتابشناختی مورد نیاز، در ابتدا پیش پردازش شامل نرمال سازی، حذف کلمات اضافی صورت گرفت. سپس دومین مرحله با فنون مختلف متن کاوی مانند تکرار واژگان و خوشه بندی انجام شد. درمرحله سوم شباهت‌سنجی و تجزیه وتحلیل دانش انجام شد.

یافته‌ها: یافته‌ها نشان داد در ۸ درصد نشریات مورد بررسی، میزان همخوانی یا تطابق محتوای مقالات با دامنه موضوعی مندرج در وب سایت به میزان ۹۵ درصد رعایت شده بود. کمترین میزان همخوانی ۴۰ درصد است که در ۲۰ درصد

[1] journals.msrt.ir

نشریات مورد بررسی مشاهده گردید. در این میان ۴۷ درصد نشریات به میزان ۶۰ درصد، ۲۳ درصد نشریات به میزان ۷۰ درصد و ۲ درصد نشریات به میزان ۸۰ درصد به موضوعات اعلامی پایبند بودند.

نتیجه‌گیری: نتایج حاصل از این ارزیابی ها می‌تواند یاریگر مسئولان بالادستی و سیاست‌گذاران عرصه علم و دانش کشور در برنامه‌ریزی های پژوهشی، ارزیابی، رصد پیشرفتهای علمی و ارتقای کیفیت فعالیتهای پژوهشی باشد. همچنین دست اندرکاران نشریات به‌منظور ارتقای محتوای نشریه و تدوین شیوه نامه های استاندارد می توانند از نتایج این ارزیابی ها بهره‌مند گردند. در پیش گرفتن رویکرد شفاف موضوعی از سوی نشریات می تواند دستاوردهای سودمندی برای جامعهٔ دانشگاهی داشته باشد. با مشخص شدن گرایش‌های موضوعی نشریه‌های علمی، هویت بارزتری برای هر نشریه ایجاد می‌شود؛ به طوری که هر متخصصی می‌تواند انتظار داشته باشد که نشریه یا نشریه های خاصی، زمینه های مطالعاتی وی را پوشش می دهد.

کلیدواژه ها: نشریه علمی، متن‌کاوی، دامنه موضوعی نشریه، تحلیل محتوا، مقالات نشریه، مشابهت‌یابی

فهرست مطالب

فصل اول

کلیات پژوهش

۱-۱ مقدمه

نشریات علمی مهم‌ترین محمل برای ارائه آخرین یافته‌های پژوهشی و فناورانه است و منبع راهبردی برای مطالعه روند رشد و تغییر پژوهش و فناوری محسوب می گردد (بهکامی و دیم[1]، ۲۰۱۲). این ابزار کارآمد اطلاع رسانی، از مشخصه های اصلی ورود یک نظام اجتماعی به دوره توسعه یافتگی تلقی می شود که در شکل دهی به شبکه پیچیده تبادل اطلاعات در سطح جهان از نقش ویژه ای برخوردار است. نشریات علمی به طور معمول نموداری از حیات علمی جامعه به شمار می روند که مورد توجه نویسندگان و پژوهشگران رشته های مختلف در سطح وسیعی است. مقاله های نشریات به دلیل تازگی و سرعت در ارائه آخرین دستاوردهای علمی، تنوع و کوتاه بودن مطالب، ارائه تحلیل های دقیق از موضوع های پیچیده علمی، دسترسی آسان، ارائه نقد و بررسی های علمی، به‌عنوان بستری برای انتشار و ارائه جدیدترین نتایج و دستاوردهای پژوهش های حوزه های مختلف علمی به شمار می وند. به سبب این ویژگی، نشریات در میان سایر مدارک و منابع علمی از جذابیت بیشتری برای نویسندگان و پژوهشگران بیشتری برخوردار هستند (رضایت،۱۳۹۴)؛ چرا که مقالات علمی به پژوهشگران کمک می‌کند دانش رشته مربوطه خود را به روز نگه دارند و روند تحقیقات و مطالعات خود را مدیریت کنند (آبل و نیولین[2]، ۲۰۰۲). بنابراین نشریات علمی و پژوهش‌های دانشگاهی همواره دارای ارتباط و مزایای متقابل برای یکدیگر هستند. از یک‌سو، یافته‌های پژوهش‌های نوین عمدتاً در مجلات علمی منتشر می‌شوند (براون[3]، ۲۰۰۵) و از سوی دیگر تعداد مخاطبان و شناخت آنان از مجلات نیز مبتنی بر پژوهش‌های با کیفیت منتشر شده در انها است. به بیان دیگر، پژوهشگران برای افزایش آگاهی و انجام پژوهش‌های جدید به مجلات علمی نیاز دارند و اشتهار مجلات علمی نیز از طریق انتشار پژوهش‌های مطرح و اثربخش، کسب می‌شود.

بر اساس مطالب پیش گفته و با توجه به نقش و اهمیتی که نشریات علمی در عرصه علم و دانش دارند، لازم است به صورت دوره ای از نظر کمی و کیفی مورد ارزیابی قرار گیرند. نتایج حاصل از این ارزیابی ها می‌تواند یاریگر مسئولان

[1] Behkami & Daim
[2] Abel & Newlin
[3] Brown

بالادستی و سیاستگذاران عرصه علم و دانش کشور در برنامه‌ریزی های پژوهشی، ارزیابی، رصد پیشرفتهای علمی و ارتقای کیفیت فعالیتهای پژوهشی باشد. همچنین دست اندرکاران نشریات بمنظور ارتقای محتوای نشریه و تدوین شیوه نامه های استاندارد می توانند از نتایج این ارزیابی ها بهره مند گردند.

این قبیل ارزیابی ها بر اساس ملاکها، معیارها و استانداردهای علمی، یکی از فرایندهای رایجی است که وزارت علوم، تحقیقات و فناوری، پایگاه استنادی علوم جهان اسلام و مسئولان نشریات به منظور بررسی نقاط قوت و ضعف نشریات مدنظر قرار دارند. به طور معمول ارزیابی ها در دو بخش ساختاری و محتوایی انجام می‌گردد. بخش ساختاری به اطلاعات شناسنامه ای و چارچوب وبگاه نشریه برمی‌گردد؛ اما با توجه به رشد روزافزون تولیدات علمی، تحلیل محتوای مقاله ها و نشریه ها، دسته بندی موضوعی آنها و شناخت انواع مطالب منتشرشده در نشریه های تخصصی برای دسترسی سریع و آسان به اطلاعات، ضروری بنظر می‌رسد.

در فرایند ارزیابی محتوایی نشریات، یکی از روشهای علمی و معتبر تحلیل محتوایی است. تحلیل محتوا، متشکل از مجموعه ای از روندهای گردآوری و سازمان دهی اطلاعات به شکلی استاندارد است که تحلیل گر را قادر می سازد دربارۀ ویژگی ها و مفاهیم موجود در نشریه، استنباط هایی داشته باشد. بدون شک برای پژوهشگران و دست اندرکاران نشریات این موضوع دارای اهمیت است که بدانند در زمینه علمی موضوعات چه روندی را طی کرده است. تحلیل محتوا، ابزاری قدرتمند برای بررسی روندها، الگوی اسناد، الگوی نوشتاری و غیره است. همچنین اهداف پژوهشگران را می توان بر اساس داده های موجود در نشریه ها و موضوع هایی که آنها برای مقاله خود انتخاب می‌کنند، استنباط نمود (ماجی، جال و مهارانا[1] ، ۲۰۱۶). در این راستا انتظار می‌رود محتوای مقالات، انعکاس دهنده پژوهش های مرتبط با دامنه و اهداف نشریه و دربردارنده اولویتها، گرایش ها و موضوع های بنیادین موجود در یک مقطع زمانی باشد. از این رو، بررسی موضوع هایی که در مقالات نشریه بدان پرداخته شده، علاوه بر اینکه تصویری عینی از شرایط موجود و انعکاس پژوهش‌ها در آن نشریه را نشان می دهد، می تواند نشان دهنده گرایش های موضوعی و کمبودهای احتمالی موجود در حوزه های مورد

[1] Maji, jal & meharata

بررسی باشد. بعبارتی در گسترهٔ وسیع منابع منتشرشده در یک حیطهٔ علمی و پژوهشی، انجام یک مطالعهٔ تطبیقی، علاوه بر آشکارسازی سیر تکوینی مطالعات آن حیطه و شناسایی تفاوتها و شباهتها، راهنمایی برای ساماندهی و مدیریت هرچه بهتر نشر آثار پژوهشی علمی به شمار میآید.

با هدف دستیابی به این مهم، گام نخست، تحلیل محتوای مقالات نشریه مبتنی بر تشریح اهداف و برنامه های آن است. به کارگیری این روش موجب استنتاج معتبر از اطلاعات موجود در متن، فراهم شدن دانش، تشریح وسیع اطلاعات، بینش جدید و راهنمای کاربردی برای عمل میشود (قهنویه و همکاران، ۱۳۹۰؛ هو و هئو[1]، ۲۰۱۳). در این روش، محتوا به صورت نظامدار و کمی توصیف میشود؛ به طوریکه داده های کیفی به داده های کمّی تبدیل میشوند. همچنین بررسی گزارش نتایج مطالعات تطبیقی به افراد متخصص در یک رشتهٔ علمی این امکان را میدهد که از دانش موجود در ارتباط با حیطهٔ پژوهشی و شغلی خود بهرهٔ لازم را ببرند و نقاط ضعف و قوت را دریابند.

به منظور انجام تحلیل محتوا، استخراج خودکار داده های متنی و استفاده از فنون متن کاوی ضروری به‌نظر می‌رسد. سالوم و همکاران[2] (۲۰۱۸) استفاده از فنون متن کاوی را روشی برای شناسایی موضوعات متون علمی و سیر تکاملی این موضوعات بیان می کنند. یکی از الگوریتمهای مهم و مفید متن کاوی، خوشه بندی است. با انجام عملیات خوشه بندی، حیطه گسترده ای از داده های پراکنده در گروه های مدون و سازمان یافته قرار میگیرند .گروههای متعدد ایجاد شده با برخورداری از ویژگیهای مشترک درون هر گروه دارای ارتباط ساختاری با یکدیگر هستند. با این روش داده های مربوط به مقالات درون خوشه های واحد قرار میگیرند به گونه ای که مقالات درون هر خوشه دارای حداکثر شباهت با یکدیگر و حداقل شباهت با دیگر خو شه ها هستند. روشهای متن کاوی و به طور خاص مدل سازی موضوعی قادر است علاوه بر کشف موضوع های پنهان و ارتباط آن با موضوع های آشکار، موضوع های کمتر شناخته شده یا کمتر پرداخته شده را شناسایی کند. تکنیک متن کاوی به عنوان یکی از روشهای تحلیل محتوا از طریق پردازش، استخراج و مرتب سازی اطلاعات به

[1] Woo H, Heo
[2] Salom et al.

ترسیم مدل مفهومی و مصورسازی اطلاعات میپردازد. همچنین امکان تحلیل، مسیریابی، نمایش و آشکارسازی ساختار و دانش مفید و ضمنی را از میان انبوهی از داد هه ای ساختارنیافته آشکار میکند (شفرین و برون[1]، ۲۰۰۴). این نوع شناسایی دانش ضمنی، اهمیت زیادی در تعیین اولویت های پژوهشی، تدوین برنامۀ راهبردی سیاست گذاران، اطلاع از شکاف موضوع‌های پژوهشی پژوهشگران ایرانی، تأثیرات بین رشته ای، ارتقای جایگاه حرفه ای و شغلی پژوهشگران و پژوهشگران هر رشته داشته و مسیری هموار را پیش روی برنامه ریزان و سیاست گذاران علمی-پژوهشی در تدوین برنامه های راهبردی قرار می دهد.

نتایج به دست آمده از مطالعات نشان داده که مدلسازی موضوعی نه تنها می‌تواند یک ابزار سودمند برای استخراج اطلاعات از داده های متنی باشد؛ بلکه نسبت به بسیاری از رویکردهای سنتی و روش‌های مبتنی بر خوشه نیز عملکرد بهتری در امر بازیابی اطلاعات دارند. با توجه به کاربرد مدلسازی موضوعی در درک ساختار موضوعی، ارتباطات بین اسناد ، گرایش های پژوهشی، دارای پتانسیل بالایی در تولید ایده های پژوهشی، تشویق به همکاری پژوهشگران و به طور کلی در حوزه سیاستگذاری علمی و پژوهشی دارد.

بر اساس مطالب پیش گفته، شناسایی اهداف و حوزه های موضوعی نشریات علمی، قلمروهای موضوعی گوناگون، دیداری‌سازی و مطالعه آنها از مسائلی است که نیازمند پژوهش است. در مجموع بر پایۀ توصیه های کمیسیون بررسی نشریات علمی کشور و نیز دفتر سیاست گذاری و برنامه ریزی امور پژوهشی وزارت علوم، تحقیقات و فناوری در مورد پایبندی محورهای موضوعات اعلام شده از سوی نشریات علمی، همۀ نشریات باید به سوی این خط مشی حرکت کنند. هدف کلی این رویکرد، فراهم ساختن بستر به گونه ای که جامعۀ دانشگاهی اعم از پژوهشگران، نویسندگان مقالات و خوانندگان بتواند هر نشریه را با موضوع یا موضوعات تخصصی شناسایی کنند (آیین نامه کمیسیون نشریات وزارت عتف، ۱۳۹۸)

[1] Shefrin & brown

از اینرو در پژوهش حاضر تلاش بر این است که با استفاده از روش تحلیل محتوای مقالات نشریات معتبر وزارت علوم در شش حوزه موضوعی سطح کلان، به تعیین میزان همخوانی موضوع یا موضوعات اعلام شده از سوی نشریات از یک سو و مقالات منتشر شده در آنها از سوی دیگر پرداخته شود.

۱-۲. بیان مسئله

آمار نشان می دهد که در حال حاضر ۳۵۰۰۰ نشریه در سراسر جهان منتشر می شود (اولریخ[1]، ۲۰۲۳) . در این میان جمهوری اسلامی ایران دارای بیش از ۲۰۰۰ نشریه علمی معتبر وزارتین است که سالانه بالغ بر ۶۰۰۰۰ مقاله را در برمی‌گیرد (اسفند ۱۴۰۱). از این‌رو، انتشار یافته‌های پژوهشی، برای به اشتراک گذاری دانش و همچنین اعتبار حرفه‌ای پژوهشگران در یک نشریه مناسب و مرتبط حائز اهمیت و از دشوارترین جنبه‌های انتشار نتایج پژوهش‌هاپژوهش‌ها است (وانگ و هو[2]، ۲۰۱۵).

انتخاب نشریه متاثر از عوامل گوناگونی است که در این میان عامل ارتباط دامنه و حوزه موضوعی نشریه با مقاله از اهمیت ویژه‌ای برخوردار است؛ زیرا ارسال مقالات حاصل از یک پژوهش، به نشریاتی که ربط کمتری با محتوا و نیاز خوانندگان دارند باعث اتلاف وقت پژوهشگر می‌شود. همچنین حوزه سردبیری نشریه نیز مدت زمانی را صرف بررسی و احتمالا ارسال مقاله به داوری نموده و اگر طی هر یک از مراحل مقاله رد شود این فرآیند باید مجددا برای نشریه‌ای دیگر تکرار گردد. نتایج مطالعات نشان می‌دهد متوسط مدت زمان تصمیم‌گیری و اعلام نظر جهت پذیرش اولیه مقاله به نویسنده، طولانی و به طور متوسط ۴۱ روز است (ویمن و اسمیتز[3]، ۲۰۱۷؛ نووین و همکاران[4]، ۲۰۱۸) که این مدت در ۳۱ درصد موارد به بالای ۶ ماه نیز می‌رسد (مالیگان، هال و رافائل[5]، ۲۰۱۳). دلیل اصلی رد مقاله در بسیاری از موارد، در مرحله بررسی اولیه توسط سردبیران و داوران، عدم تطابق و تناسب موضوعی مقاله با دامنه موضوعی نشریه است. در برخی موارد، نرخ رد اولیه

[1] Ulrichs
[2] Wang & Hou
[3] Huisman and Smits
[4] Nguyen et al.
[5] Mulligan, Hall, Raphael

حتی قبل از ارسال به داوری تا ۸۸ درصد اعلام شده است (اندرسون[1]، ۲۰۱۲). بنابراین ارسال مقاله به نشریه نامرتبط، علاوه بر اینکه مزایای کمتری برای ارتقای فردی و سازمانی پژوهشگر به همراه دارد منجر به از دست رفتن زمان قابل ملاحظه، تازگی، اهمیت و تاثیرگذاری یافته‌های پژوهش‌هاپژوهش‌های خواهد شد (هو، وو، لین[2]، ۲۰۱۶). این چالش با شفاف سازی حوزه های موضوعی مورد تایید نشریه جهت دریافت و پذیرش مقاله تا حد قابل توجهی بهبود خواهد یافت. همچنین تخصصی شدن نشریه های علمی موجب تخصصی شدن فرایند داوری، ویراستاری و نشر مقالات خواهد شد. از این نظر، انسجام و یکدستی بهتری در نشریه ها به وجود خواهد آمد. این وضعیت می تواند از دریافت و داوری مقاله های تکراری در دفاتر مجلات و نیز بروز سرقت علمی جلوگیری نماید.

در همین راستا و به‌منظور پاسخ به پرسش‌های پژوهش طرح حاضر بر پایه فنون متن کاوی به بررسی میزان همخوانی دامنه و اهداف موضوعی نشریات فارسی وزارت عتف با محتوای مقالات منتشر شده در انها پرداخته است. روش‌های متن‌کاوی به‌منظور تحلیل خودکار انتشارات علمی به‌کاررفته است که به بررسی اسناد به‌منظور شناسایی مضامین یا موضوعات آن‌ها می‌پردازد. از نتایج حاصل می‌توان به‌منظور تحلیل چگونگی ارتباط مباحث با یکدیگر و چگونگی تکامل آن‌ها باگذشت زمان استفاده کرد.

۱-۳. اهمیت پژوهش و ضرورت

– در ایران ۱۴۵۰ عنوان نشریه علمی معتبر توسط وزارت علوم، تحقیقات و فناوری در راستای افزایش مشارکت علمی و انتشار نتایج مطالعات و پژوهش‌هاپژوهش‌ها در سطح ملی و بین المللی منتشر می شود (اسفند ۱۴۰۱). این نشریات به لحاظ کمی و کیفی در سطوح الف تا دال رتبه بندی می شوند. در این میان ۸۵ درصد این نشریات از لحاظ درجه علمی دارای رتبه های الف و ب هستند که این امر نشان از اعتبار علمی بالای آنها دارد. افزایش تعداد نشریه های علمی ایجاب می کند که برای حفظ و توسعه کیفی آنها تلاش شود. از اینرو ارزیابی دقیق موضوعی مقالات این نشریات می تواند شاخصی از روند و گرایش

[1] Anderson
[2] Hou, Wu and Lin

فعالیتهای پژوهشی را نشان دهد. این پژوهش در جهت پایش وضعیت موجود و بهبود جایگاه نشریات کشور در حوزه مربوطه و تعیین الگوی موضوعی است و در مجموع این نکته که نشریات به چه میزان در اعلام و رعایت دامنه و حوزه های موضوعی که در آن مقاله می پذیرند شفاف عمل کرده اند قابل بررسی و حائز اهمیت است.

- در پیش گرفتن این رویکرد از سوی نشریات می تواند دستاوردهای سودمندی برای جامعهٔ دانشگاهی داشته باشد. با تخصصی شدن گرایش های موضوعی نشریه های علمی، هویت بارزتری برای هر نشریه ایجاد می شود؛ به طوری که هر متخصصی می‌تواند انتظار داشته باشد که نشریه یا نشریه های خاصی، زمینه های مطالعاتی وی را پوشش خواهد داد. تخصص گرایی فرآیندی علمی و عملی است که می توان به یاری آن، پژوهش های اعضای هیات علمی و پژوهشگران کشور را معطوف به حوزه های تخصصی کرد. در تعریف حوزه تخصص باید گفت که حوزه تخصص به آن حوزه یا گرایش خاصی اطلاق می گردد که مشتمل بر مجموعه مسائل نظام مندی است که نقشه راه فعالیت نشریه را در بازه زمانی طولانی تعیین می نماید و نتیجه آن به فراوری دانش انباشته و رسیدن به نظریه های جدید در آن حوزه علمی منتهی می گردد. بدین ترتیب، نشریه و نویسندگان آن به مرجعی در جامعه علمی تبدیل می گردند (فتاحی، ۱۳۹۴).

- با گسترش علوم و تخصصی شدن هر حوزهٔ علمی، جایگاه تخصص‌ها نیز اهمیت بیشتری می‌یابد. گرایش‌های تخصصی در رشته‌های دانشگاهی و رویکردهای کاملاً تخصصی در پژوهش، حکایت از روند روبه توسعه علم دارند. از سوی دیگر، توسعهٔ گرایش‌های تخصصی در هر کشور نشان‌دهندهٔ تلاش اندیشمندان و مدیران آن کشور در جهت توسعهٔ علم و درنوردیدن مرزهای موجود است. به بیان دیگر، تخصص گرایی، شاخص پیشرفت علم در هر جامعه تلقی می‌شود. در این میان، دانشگاه‌ها و مراکز آموزشی و پژوهشی نقش اصلی و تعیین‌کننده دارند. توسعهٔ علم و ایجاد رشته‌های جدید دانشگاهی، بویژه در مقاطع کارشناسی ارشد و دکترا عامل مهمی در توسعهٔ تخصص‌ها به‌شمار می‌روند. در همین راستا، انتظار می‌رود اعضای هیئت علمی و دانشجویان مقاطع تحصیلات تکمیلی عمدهٔ تلاش‌های خود را در بعد آموزش، پژوهش و فعالیت‌های علمی بر یک زمینهٔ تخصصیِ محدود متمرکز و در همان زمینه رشد و متخصص شوند. این روند سال‌ها در دانشگاه‌ها و مراکز آموزشی و پژوهشی جوامع پیشرفته اعمال شده و در حال حاضر وضعیت به گونه‌ای است که هر مؤسسه یا هر فرد در یک یا حداکثر دو

زمینهٔ خاص و محدود، به فعالیت اشتغال دارد. به همین دلیل، فعالیت‌های پژوهشی و تولید علم در آن جوامع از پیشرفت قابل توجهی برخوردار است. هر پژوهشگر در طول حیات علمی خود تنها یک یا دو زمینهٔ خاص را دنبال می‌کند و در همان زمینه(ها) نیز متخصص و صاحب‌نظر می‌شود. تولیدات علمی (کتاب‌ها و مقاله‌ها) نیز کاملاً تخصصی است. به عبارت دیگر، کمتر مشاهده می‌کنیم که پژوهشگران آن جوامع دچار پراکنده‌کاری و گسیختگی علمی شوند. در مقابل، نگاهی به مقاله‌ها و کتاب‌های تألیفی یا ترجمه‌ای اعضای هیئت علمی و پژوهشگران ایرانی حاکی از آن است که بسیاری از پژوهشگران دچار نوعی پراکنده‌کاری و فعالیت در هر زمینه‌ای هستند. عمومیت این وضعیت در رشته‌های علوم انسانی و اجتماعی بیش از دیگر علوم است. بنظر میرسد چنانچه هر نشریه در حوزه ای تخصصی مربوطه به جنبه های نظری و مفهومی، در عرصه کار علمی طی سال های فعالیتش متمرکز شود، خروجی های قابل استفاده و عمیق تری خواهد داشت. چنین راهبردی به عمق و غنای کارهای علمی پژوهشگران بسیار خواهد افزود و باعث می شود که پراکنده کاری و سطحی نگری به حداقل ممکن کاهش یابد (گاهنامه عتف، ۱۳۹۵). از این رو، مراجعهٔ منظم به نشریه های معین به شکل چاپی یا الکترونیکی، وی را در جریان روند پژوهش های تخصصی حوزهٔ مورد علاقه او قرار خواهد داد.

– با توجه به اهمیت انتشار یافته های پژوهشی و هزینه و زمانی که صرف انجام آن می شود؛ توجه به موضوع پژوهشهای انجام شده در دوره های زمانی مختلف حائز اهمیت است. چنین بررسی هایی نشان می دهد که در دوره های زمانی مختلف به چه موضوعاتی بیشتر یا کمتر پرداخته شده است. هنگامی که کاستی های پژوهش در موضوعاتی خاص تعیین شد می توان انتظار داشت که خط مشی پژوهشهای آینده طوری برنامه ریزی شود که کاستی ها بمنظور شناسایی اولویتهای پژوهشی انجام و تصویری کلی از سیر موضوعی که نشریات در پیش گرفته اند مشخص گردد.

– با مشخص شدن دامنه موضوعی نشریات، پژوهشگران می توانند در انتخاب و استفاده از منابع مرتبط در یک حوزه تخصصی بهتر عمل کنند؛ لذا نتایج این پژوهش می تواند در این امر نیز به کار گرفته شود. همچنین دست اندرکاران نشریات نیز می توانند در تامین نیاز جامعه هدف مخاطبان هدف دقیق تر عمل کنند.

– همچنین عمق بخشیدن به پژوهش‌ها و توانمندسازی نشریات در جهت شناسایی مرزهای علمی تخصصی متقن و معتبر،

زمینه سازی برای تعمق، ارتقا و شکوفایی خلاقیت و ابتکار علمی محققین، افزایش کارایی نشریات در جهت گره گشایی و حل مسائل واقعی کشور، امکان ایجاد شبکه های علمی تخصصی بین نویسندگان، استفاده از اطلاعات تخصصی در فعالیت تصمیم گیری مدیران ارشد مراکز پژوهشی وزارت علوم، تحقیقات و فناوری از دیگر نتایج طرح حاضر در رفع برخی چالشها خواهد بود.

۱-۴. هدف پژوهش

هدف اصلی پژوهش حاضر، تحلیل میزان همخوانی مقالات منتشر شده با دامنه موضوعی نشریات وزارت علوم، تحقیقات و فناوری بر پایه فنون متن‌کاوی است.

در این راستا اهداف جزئی به شرح زیر تعریف می گردد:

- بررسی میزان همخوانی محتوای مقالات نشریات وزارت علوم در شش حوزه موضوعی علوم انسانی، فنی و مهندسی، علوم پایه، هنر و معماری، کشاورزی و منابع طبیعی و دامپزشکی
- فراوانی نشریات بر اساس میزان همخوانی محتوای مقالات نشریه با حوزه موضوعی مندرج در وبگاه به تفکیک حوزه موضوعی

۲-۴. تعریف‌های مفهومی و عملیاتی

- متن‌کاوی
  - تعریف مفهومی

متن‌کاوی تکنیکی میان‌رشته‌ای است و این امکان را فراهم می‌کند تا بتوان از طریق شناسایی و اکتشاف الگوهایی در داده‌های متنی از کلان داده‌ها به شکلی مفید استفاده نمود (تروینس و همکاران[1]، ۲۰۱۴). همچنین متن‌کاوی می‌تواند در جهت درک

[1] Truyens et al.

بهتر اطلاعات موجود در اسناد به کار گرفته شود. درواقع متن‌کاوی امکان بازیابی و درک اطلاعات پنهان در متن‌ها را فراهم می‌کند و برای کشف ساختار، الگوها و دانش در مجموعه‌های متنی بزرگ به کار می‌رود (تسنگ[1] و همکاران، ۲۰۰۷).

- تعریف عملیاتی

در این پژوهش از متن‌کاوی جهت شناسایی و اکتشاف محتوای موضوعی مقالات نشریات استفاده گردید. بدین ترتیب که ابتدا با مراجعه به وبگاه هر نشریه اهداف و دامنه موضوعی و همچنین اطلاعات کتابشناختی مقالات بازه زمانی مورد بررسی استخراج و با استفاده از فنون متن کاوی به داده هایی که قابل خواندن و پردازش برای ماشین باشد تبدیل شد.

- **خوشه‌بندی موضوعی**
  - تعریف مفهومی

یکی از الگوریتمهای مهم و مفید متن کاوی، خوشه بندی است. با انجام عملیات خوشه بندی، حیطه گسترده ای از داده های پراکنده در گروه های مدون و سازمان یافته قرار میگیرند .گروههای متعدد ایجاد شده با برخورداری از ویژگیهای مشترک درون هر گروه دارای ارتباط ساختاری با یکدیگر هستند. با این روش داده های مربوط به مقالات درون خوشه های واحد قرار میگیرند به گونه ای که مقالات درون هر خوشه دارای حداکثر شباهت با یکدیگر و حداقل شباهت با دیگر خو شه ها هستند (لامبا و مدهوسدان[2]، ۲۰۱۸).

- تعریف عملیاتی

در این پژوهش محتوای نشریات با استفاده از الگوریتم فضای برداری مورد بررسی و تحلیل قرار گرفته است. بدین ترتیب که برای هر مدرک یک بردار مبتنی بر فرکانس تکرار واژه ها تشکلیل و سپس با برداری که بر اساس واژه های کلیدی مستخرج از دامنه و حوزه های موضوعی ایجاد گردیده شباهت سنجی شد.

[1] Tseng et al.
[2] Lamba & Madhusudhan

فصل دوم

مبانی نظری

و

مرور پیشینه‌های پژوهش

۱-۲. مقدمه

در این فصل مبانی نظری تبیین گردیده و پیشینه‌های پژوهش مرور شده است. به بیان دیگر، ابتدا به مبانی نظری پژوهش در مباحث مرتبط با داده‌کاوی و متن‌کاوی پرداخته‌شده و سپس پیشینه‌های پژوهش‌های در سطح ملی و بین‌المللی مرور شده و در پایان به استنتاج و نتیجه‌گیری از مرور پیشینه‌های پژوهش پرداخته شده است.

۲-۲. مبانی نظری

۱-۲-۲. داده‌کاوی

پیشرفت‌های به وجود آمده در جمع‌آوری داده‌ها و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از داده‌ها روبرو شویم. داده‌کاوی کوششی برای به‌دست‌آوردن اطلاعات مفید از میان این داده‌هاست و رشد بی‌رویه داده‌ها در سطح بین‌المللی اهمیت داده‌کاوی را دوچندان کرده است. فناوری مدیریت پایگاه داده‌های پیشرفته انواع مختلفی از داده‌ها را می‌تواند در خود جای دهد، درنتیجه تکنیک‌های آماری و ابزار مدیریت سنتی برای تحلیل این داده‌ها کافی نیست و استخراج دانش از این حجم بسیار زیاد داده‌ها چالشی اساسی تلقی می‌شود.

برای داده‌کاوی تعاریف متعددی وجود دارد. برخی از این تعاریف عبارت‌اند از: فرآیند به خدمت گرفتن یک روش‌شناسی رایانه‌ای که با استفاده از فنون مختلف، دانش را به طور مستقیم از داده‌ها استخراج می‌کند (بارس و کمپور[1]، ۲۰۰۸). داده‌کاوی جستجویی است برای اطلاعات جدید و نوین از میان مقادیر کلان داده‌ها و فرآیندی است مشارکتی میان انسان و کامپیوتر (یوکسل‌ترک و همکاران[2]، ۲۰۱۴). داده‌کاوی فرآیند اکتشاف و تحلیل داده‌ها به‌وسیله ابزار خودکار و نیمه‌خودکار به‌منظور اکتشاف الگوهای معنی‌دار و قواعد است (پرانتا و اسکینر[3]، ۲۰۱۵).

اصولاً داده‌کاوی، پایگاه‌های داده‌ی بزرگ را به‌عنوان منبع بالقوه‌ای از دانش ارزشمند برای تصمیم‌گیری در نظر می‌گیرد. در فرآیند داده‌کاوی بهترین نتیجه زمانی حاصل می‌شود که دانش یک فرد خبره در خصوص یک مسئله با توانایی‌های کامپیوتر ترکیب شود (خان[4]، ۲۰۱۶). فقط در این صورت است که می‌توان بدون نفی توانایی‌های فرد خبره، سیستم‌های کامپیوتری را در خدمت نیرومندتر ساختن او قرارداد.

۱-۱-۲-۲. اهداف داده‌کاوی

دو هدف اساسی داده‌کاوی، پیش‌بینی[5] و توصیف[6] است. در عملیات پیش‌بینی بعضی از متغیرها یا حوزه‌هایی از مجموعه‌های داده به‌منظور پیش‌بینی ارزش ناشناخته یا ارزش آینده داده‌های دیگر مورداستفاده قرار می‌گیرد. همچنین داده‌کاوی بر یافتن الگوهای تشریحی داده‌ها که به‌وسیله انسان می‌تواند تعبیر شود، تمرکز می‌کند. درنتیجه داده‌کاوی را می‌توان دریکی از دو گروه زیر جای داد:

-**پیش‌بینی**: این روش با استفاده از مجموعه داده‌ها، مدل‌هایی را برای توضیح سیستم تولید می‌کند که با استفاده از آن‌ها می‌توان عملکرد متغیرهای مختلف را پیش‌بینی کرد.

---

[1] Baars and Kemper
[2] Yukselturk et al.
[3] Pranata and Skinner
[4] Khan
[5] Prediction
[6] Description

- **توصیف**: اطلاعات جدید را بر اساس مجموعه داده‌های در دسترس تولید می‌کند این داده‌ها الگوهای رفتاری متغیرها را تشریح می‌کند.

هدف از داده‌کاوی پیش‌بینی‌کننده تولید مدلی است که با استفاده از یک کد اجرایی وظایفی چون پیش‌بینی، دسته‌بندی، تخمین مقدار، تخمین عملکرد و برخی دیگر از موارد را انجام دهد. هدف از داده‌کاوی تشریحی دستیابی به درکی کامل از سیستم تحت بررسی با استفاده از الگوهای پنهان در آن و روابط درون مجموعه‌های داده است(کانتاردزیک[1]، ۲۰۱۱).

۲-۱-۲-۲. ریشه‌های داده‌کاوی

پایه‌های اصلی داده‌کاوی بر دو اصل آمار و یادگیری ماشین[2] استوار است. آمار نیز ریشه در ریاضیات و منطق داشته و بنابراین داده‌کاوی نیز علاوه بر آمار ریشه در این دو علم دارد. در مقابل یادگیری ماشینی نیز علمی مبتنی بر رایانه است که اصول آن را در هوش مصنوعی می‌توان یافت. تضادی که در اینجا آشکار می‌شود این است که آمار به دلیل طبیعت ریاضی خود متمایل به فرموله کردن مسائل و مدل‌سازی است، درحالی‌که یادگیری ماشینی مسائل را با استفاده از الگوریتم‌ها حل می‌کند. اینجاست که باید نسبت به ترکیب این دو علم برای استفاده از آن‌ها در داده‌کاوی اقدام کرد.

داده‌کاوی رویه‌های تحلیلی را در زمینه‌های آمار، ریاضیات، تجارت و نظریه اقتصاد پیوند می‌زند. داده‌کاوی علاوه بر علوم فوق به‌خاطر استفاده از اصول اساسی مدل‌سازی از نظریه کنترل نیز استفاده می‌کند. این نظریه عموماً در سیستم‌های مهندسی و فرآیندهای صنعتی مورداستفاده قرار می‌گیرد. بنابراین داده‌کاوی یک فناوری میان‌رشته‌ای است و برای استفاده مؤثر از آن باید از علوم تشکیل‌دهنده آن شناخت کافی داشت. البته زمانی که بخواهیم از داده‌کاوی برای مقاصد نوآورانه و خلاقانه‌تر استفاده کنیم، نیاز به این شناخت به‌مراتب عمیق‌تر می‌شود.

[1] Kantardzic
[2] Machine Learning

علی‌رغم ارتباط میان داده‌کاوی و آمار، تفاوت‌های اساسی میان این دو علم وجود دارد. آمار، علمی تأییدی[1] است؛ یعنی کوشش دارد مفروضاتی را با استفاده از فنون مختلف تصدیق یا رد کند، درحالی‌که داده‌کاوی یک علم اکتشافی[2] است، بدین معنی که سعی دارد الگوهای دانشی داده‌های موجود را کشف نماید. همچنین آمار استنتاجی از نمونه‌های کوچک و تعمیم آن‌ها به جامعه استفاده می‌کند و ماهیتاً توان پردازش نمونه‌های بزرگ را ندارد درحالی‌که در داده‌کاوی از نمونه‌های بسیار بزرگ و حتی خود جامعه استفاده می‌شود. دلیل این امر استفاده این فناوری از روش‌های پیشرفته کامپیوتری است که توان پردازش بالایی دارد و نهایتاً آمار فقط می‌تواند نمونه را به جامعه‌ای که از آن انتخاب‌شده تعمیم دهد، درحالی‌که در داده‌کاوی تحلیل‌ها بر روی کل جامعه انجام می‌شود (جهرمی و همکاران، ۲۰۱۶).

۳-۱-۲-۲. کاربردهای داده‌کاوی

تاریخچه کشف دانش در پایگاه‌های اطلاعاتی که امروزه به داده‌کاوی مشهور است، قدمت چندانی ندارد. در اوایل دهه ۱۹۹۱، هنگامی‌که اصطلاح کشف دانش در پایگاه‌های اطلاعاتی برای نخستین بار مطرح شد، هجمه‌ای همگانی به سمت طراحی الگوریتم‌های داده‌کاوی صورت پذیرفت (ماربان و همکاران[3]، ۲۰۰۹) و این‌زمانی بود که شرکت‌ها اقدام به ذخیره‌سازی مقادیر عظیم داده‌ها کردند و به دنبال روش‌هایی برای استفاده از این انبار داده‌ها بودند (نیکلسن[4]، ۲۰۰۳).

باتوجه‌به توان تحلیل بالای فناوری داده‌کاوی و باوجود قدرت پردازش بی‌نظیر آن، از این فناوری می‌توان برای حل مسائل بی‌شماری در دنیای واقعی استفاده کرد. برخی از کاربردهای داده‌کاوی عبارت‌اند از:

- تحلیل رفتار افراد و گروه‌ها (ماربان و همکاران، ۲۰۰۹)؛
- پردازش اطلاعات پزشکی (مسروپیان و اوسیانیکوف[5]، ۲۰۱۴)؛

[1] Confirmatory
[2] Exploratory
[3] Marbán et al.
[4] Nicholson
[5] Mesropyan & Ovsyannikov

- تشخیص الگوهای رفتاری مصرف‌کنندگان (کانتاردزیک[1]، ۲۰۱۱)؛

- یافتن پروتئین‌های مختلف از نقشه ژنی[2] موجودات زنده (میترا و آچاریا[3]، ۲۰۰۳)؛

- هوشمندی کسب‌وکار و کاهش ابهامات ناشی از محیط (ماارییناکوس و داسکالکی[4]، ۲۰۱۶)؛

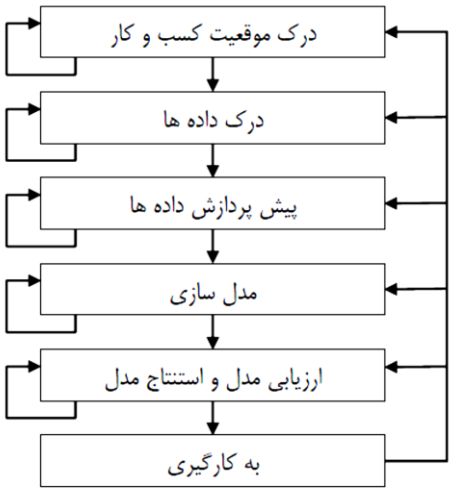
- مبارزه با جرم و جنایت و تشخیص الگوهای رفتاری گروه‌های تروریستی (عبدالله و همکاران[5]، ۲۰۱۶) و

- بهینه‌سازی تصمیمات و تخمین‌ها در بازارهای مالی (جین و همکاران[6]، ۲۰۱۳).

۴-۱-۲-۲. فرآیند داده‌کاوی

باتوجه به این امر که داده‌کاوی فرآیند اکتشاف مدل‌های گوناگون، خلاصه‌ها و ارزش‌های نشأت‌گرفته از مجموعه خاصی از داده‌ها است، برای پیاده‌سازی چنین فرآیندی باید از روش‌شناسی خاصی استفاده کرد. در این راستا روش‌شناسی فرآیند استاندارد میان صنعتی داده‌کاوی[7] به‌وسیله تحلیل نمایندگی‌های دایملر کرایسلر ایجاد شد (چاپمن[8] و همکاران، ۲۰۰۰) این روش‌شناسی توانمند و منعطف جهت ارتقای شایستگی داده‌کاوی در حل مسائل سازمانی است. بر اساس این روش، یک پروژه داده‌کاوی مبتنی بر چرخه عمر، متشکل از شش گام است و این گام‌ها به‌صورت مستمر و تکراری در تمام فرآیند داده‌کاوی به کار گرفته می‌شود. گام‌های روش‌شناسی داده‌کاوی CRISP به شرح زیر است (چاپمن و همکاران، ۲۰۰۰؛ لیو و یو[9]، ۲۰۱۵):

[1] kantardzic
[2] Genomic Map
[3] Mitra and Acharya
[4] Marinakos & Daskalaki
[5] Abdallah et al.
[6] Jain et al.
[7] Cross Industry Standard Process(CRISP)
[8] Chapman
[9] Liu and Yu



نمودار ۱-۲. مراحل داده‌کاوی CRISP

۲-۲-۲. متن‌کاوی

متن‌کاوی اکتشاف و استخراج دانش جذاب[۱] و غیر بدیهی[۲] از متن آزاد یا غیر ساختارمند است و تمامی فعالیت‌هایی که به‌نوعی به دنبال استخراج دانش از متن هستند را شامل می‌شود (هانگ و ژانگ[۳]، ۲۰۱۲؛ لی[۴] و همکاران، ۲۰۰۸). هدف از متن‌کاوی کشف اطلاعات جدید و ناشناخته، به‌وسیله استخراج خودکار اطلاعات از داده‌های متنی بدون ساختار یا نیمه ساختاریافته است و کاربردهای گسترده‌ای در تجزیه‌وتحلیل و پردازش مدارک متنی دارد (بلی[۵]، ۲۰۱۲؛ آبرامسون[۶] و همکاران، ۲۰۱۴؛ وانک، بلی و هکرمن[۷]، ۲۰۱۲؛ استیورز و گریفت[۸]، ۲۰۰۷؛ جلودار[۹] و همکاران، ۲۰۱۹).

[1] Interesting
[2] Non-trivial
[3] Hung & Zhang
[4] Lee
[5] Blei
[6] Abramson
[7] Wang, Blei & Heckerman
[8] Steyvers & Griffiths
[9] Jelodar

پیشرفت پیوسته فناوری باعث افزایش شدید حجم اطلاعات، خصوصاً اطلاعات علمی و تکنیکی شده است. درنتیجه، پیگیری اطلاعات به وظیفه‌ای چالش‌برانگیز برای دانشمندان تبدیل شده است. بنابراین، تکنیک‌های متن‌کاوی برای کمک به دانشمندان در زمینهٔ کسب اطلاعات سودمند از حجم بسیار زیادی از اطلاعات، موردنیاز است. متن‌کاوی و یا کشف دانش از متن اشاره به فرآیندی است که موجب به‌دست‌آوردن الگوهای غیر بدیهی، جالب و باکیفیت بالا و همچنین اطلاعات و دانش از اسناد متنی ساختار نیافته می‌شود. متن‌کاوی (که به‌عنوان داده‌کاوی متنی[1] نیز شناخته می‌شود) به جست‌وجو در میان داده‌های متنی برای استخراج اطلاعات مفید می‌پردازد که معمولاً طبیعتی ساختار نیافته دارد (الویدیان[2] و همکاران، ۲۰۱۵). هدف اولیه متن‌کاوی، بازیابی اطلاعات از متون ساختار نیافته و همچنین ارائه دانش به‌صورت خالص و چکیده برای کاربران است (آنانیادو و مکنوت[3]، ۲۰۰۶). هدف از متن‌کاوی قادر ساختن کاربران برای جمع‌آوری، ذخیره، کشف، و تفسیر دانش موردنیاز برای پژوهش و آموزش مؤثر و نظام‌مند است. متن‌کاوی شامل سه فعالیت اصلی است: بازیابی اطلاعات که بازیابی متون مربوط به پرسش استفاده‌کنندگان است؛ خلاصه اطلاعات که شناختن و استخراج نکات ریز متون که مربوط به پرسش است؛ داده‌کاوی رابطه مستقیم یا غیرمستقیم بین قسمت‌های اطلاعات استخراجی از متون را پیدا می‌کند. به بیان دیگر متن‌کاوی شاخه‌ای از داده‌کاوی یا همان کشف دانش است (توماس، مک‌نات و آنانیادو[4]، ۲۰۱۱).

هدف متن‌کاوی، کشف اطلاعات ناشناخته‌ای است که هنوز کسی نمی‌داند و بنابراین مستند نشده است (گولاب[5]، ۲۰۰۶). ازآنجاکه بسیاری از اطلاعات به شکل متن ذخیره‌شده‌اند، متن‌کاوی، ارزش اقتصادی بسیار بالایی در پی خواهد داشت. دانش ممکن است از منابع گوناگون اطلاعاتی به‌دست‌آمده باشد، اما متون ساختار نیافته، بیشترین منابع دانش در دسترس را تشکیل می‌دهد. مسئله کشف دانش از متون، استخراج مفاهیم صریح و غیرصریح و روابط معنایی میان مفاهیم با استفاده از فنون پردازش زبان طبیعی است. هدف استخراج دانش، به دست آوردن آگاهی درباره کلان داده‌های متنی است. کشف دانش از

[1] Text Data Mining
[2] Alwidian
[3] Ananiadou and McNaught
[4] Thomas, McNaught & Ananiadou
[5] Golub

متن ریشه در پردازش زبان طبیعی دارد؛ اما از روش‌های آماری، یادگیری ماشینی، استدلال استخراج اطلاعات، مدیریت دانش و دیگر رشته‌های مرتبط برای فرآیند کشف، وام گرفته شده است. کشف دانش از متن، نقش مهمی در ظهور برنامه‌هایی مانند فهم متن ایفاء می‌کند. متن‌کاوی با استفاده از پردازش زبان طبیعی، کاربردپذیری کشف دانش از داده‌ها را به‌طور قابل توجهی افزایش داده است. این بدان معناست که نیازی نیست فرآیند کشف دانش از داده‌ها را تنها به آن دسته از اطلاعات موجود در پایگاه‌های ساختاریافته محدود کنیم.

باتوجه به اینکه بیشتر اطلاعات ارزشمند برای استخراج هم‌اکنون در متون زبان طبیعی وجود دارد، پردازش زبان طبیعی می‌تواند فنون موردنیاز برای متن‌کاوی را فراهم کرده و دانش را به‌طور خودکار از این متون استخراج کند. متن‌کاوی درباره جستجوی انگاره‌ها در متن زبان طبیعی است و عبارت است از فرآیند تحلیل متن به‌منظور استخراج اطلاعات از حجم عظیمی از متون غیر ساختاریافته. متن‌کاوی نوعی فناوری است که امکان کشف انگاره‌ها و گرایش‌ها را به‌طور نسبتاً خودکار از درون‌متن غیر ساختاریافته آزاد فراهم می‌کند. همان‌طور که گفته شد، متن‌کاوی به‌عنوان امتداد طبیعی داده‌کاوی به شمار می‌رود و استفاده از همان فناوری‌های داده‌کاوی در قلمرو اطلاعات متنی است. متن‌کاوی در مقایسه با داده‌کاوی فرآیند پیچیده‌تری است، زیرا با آن دسته از داده‌های متنی سروکار دارد که غیر ساختاریافته هستند. متن‌کاوی قلمرویی چند رشته‌ای است و شامل بازیابی اطلاعات، تحلیل متن، دسته‌بندی اطلاعات، طبقه‌بندی اطلاعات، دیداری‌سازی اطلاعات، فناوری پایگاه‌های اطلاعاتی، یادگیری ماشینی و داده‌کاوی است. علاوه بر این از ابزارهای متن‌کاوی برای کمک به پژوهشگران در زمینه‌های مختلف استفاده می‌شود. به‌عنوان‌مثال، یک اخترشناس که علایم اشعه ایکس را در ناحیه‌ای از فضا کشف می‌کند ممکن است تمایل داشته باشد که در پیشینه‌های و منابع الکترونیکی جستجو نماید تا ببیند که آیا تاکنون هیچ نوع اشعه مادون قرمزی در همان ناحیه کشف‌شده است یا خیر؟ یک زیست‌شناس که دارای فهرستی از تعداد ۱۹۹ ژن است که در مقالات مختلف شناسایی‌شده‌اند ممکن است تمایل داشته باشد که به سراغ پژوهش‌های منتشرشده موجود برود و در جستجوی مقالاتی باشد که درباره عملکرد این ژن‌ها توضیح داده‌اند (شارما[1]، ۲۰۰۵).

[1] Sharma

متن‌کاوی که به تحلیل هوشمند متن، داده‌کاوی متنی یا کشف دانش در متن، نیز مشهور است عموماً به فرآیند استخراج دانش و اطلاعات موردعلاقه و مهم از مجموعه متنی غیر ساختاریافته اشاره دارد. به‌عبارت‌دیگر، متن‌کاوی فرآیند تحلیل طبیعی متن به‌منظور کشف و ثبت اطلاعات معنایی برای درونداد و ذخیره‌سازی در یک ساختار سازمان‌یافته دانش است. یکی از تکنیک‌های مفید متن‌کاوی دسته‌بندی است که دسته‌بندی برای کشف توزیع‌ها و انگاره‌های داده‌های موردعلاقه در کلان داده‌ها مورداستفاده قرار می‌گیرد. با استفاده از این تکنیک می‌توان بدون اتکاء بر هیچ دانش پیش‌زمینه‌ای، ساختارها یا دسته‌های موردعلاقه را به‌طور مستقیم از داده‌ها شناسایی نمود (پونر - پوراتا[1] و همکاران، ۲۰۰۷).

از مهم‌ترین مزایای متن‌کاوی می‌توان به موارد زیر اشاره نمود:

۱- استخراج اطلاعات مفید و سودمند از کلان داده‌ها در زمان کوتاه؛

۲- پیش‌بینی جنبه‌های آینده بر اساس مهیاکردن آمار و ارقام و مشاهدات؛

۳- طراحی و ارائه الگوهایی از داده‌های در دست بررسی؛

۴- مشاهده و تحلیل اطلاعات متنی سازمان‌های امنیتی در منابع اینترنتی و شبکه‌های اجتماعی؛

۵- تحلیل، ذخیره‌سازی و دسترس‌پذیری اطلاعات وب‌سایت‌ها و موتورهای جست‌وجو به‌منظور پردازش و جست‌وجوی مؤثرتر و با دقت‌تر

۶- تحلیل لغوی و تشخیص الگوها به‌منظور مطالعه توزیع فرکانس واژگان (هاشمی[2] و همکاران، ۲۰۱۵؛ ربهولز-شومان[3] و همکاران، ۲۰۱۲؛ راجمان و وسلی[4]، ۲۰۰۴؛ سالوم[5] و همکاران، ۲۰۱۷الف؛ اوکالهان[6] و همکاران، ۲۰۱۵).

۱-۲-۲-۲. رابطه متن‌کاوی و داده‌کاوی

[1] Pons-Porrata
[2] Hashimi
[3] Rebholz-Schuhmann
[4] Rajman and Vesely
[5] Salloum
[6] O'callaghan

متن‌کاوی شامل طیف گسترده‌ای از وظایف است که می‌تواند اطلاعاتی را در مورد جنبه‌های مختلف متون به ارمغان آورد. وظایف معمول استخراج متن شامل موارد زیر است: (ماینر[1] و همکاران، ۲۰۱۲):

- طبقه‌بندی اسناد[2] - اختصاص سند به یک یا چند طبقه از پیش تعریف‌شده (به‌عنوان‌مثال، اختصاص مقاله روزنامه به یک یا چند طبقه، برچسب‌گذاری نامه‌های الکترونیکی به‌عنوان هرزنامه)؛

- خوشه‌بندی[3] - گروه‌بندی اسناد با توجه به شباهت آن‌ها، به‌عنوان‌مثال، به‌منظور شناسایی اسنادی که دارای یک موضوع مشترک هستند؛

- جمع‌بندی[4] (خلاصه‌سازی)- یافتن مهم‌ترین قسمت‌ها در یک یا چند سند و ایجاد متنی که به‌طور قابل‌توجهی کوتاه‌تر از سند اصلی باشد؛

- بازیابی اطلاعات[5] - بازیابی اسنادی که با کوئری مطابقت دارند و اطلاعات موردنیاز را از مجموعه بزرگی از اسناد نشان می‌دهد؛

- استخراج معنی[6] - اسناد یا قسمت‌هایی از آن را با شناسایی موضوعات پنهان، تجزیه‌وتحلیل احساسات، عقاید نشان می‌دهد؛

- استخراج اطلاعات[7] - استخراج اطلاعات ساختار یافته مانند موجودیت‌ها، رویدادها یا روابط از متون بدون ساختار؛

- استخراج روابط[8] - یافتن ارتباط بین مفاهیم یا اصطلاحات در متون؛

- تحلیل روند[9] - بررسی چگونگی تغییر مفاهیم موجود در اسناد در گذر زمان؛

- ترجمه ماشینی[10] - تبدیل متن نوشته‌شده از یک‌زبان به زبان دیگر.

---

[1] Miner
[2] categorization of documents
[3] clustering
[4] summarization
[5] information retrieval
[6] extracting the meaning
[7] information extraction
[8] association mining
[9] trend analysis
[10] machine translation

داده‌کاوی فرآیند خودکار یا نیمه‌خودکار یافتن دانش ضمنی، قبلاً ناشناخته و بالقوه مفید در مجموعه داده‌های ذخیره‌شده الکترونیکی است. دانش دارای شکلی از الگوهای ساختاری در داده‌ها است که می‌تواند برای پیش‌بینی یا ارائه پاسخ در آینده نیز مورداستفاده قرار گیرد (ویتن و فرانک[1]، ۲۰۰۲). داده‌کاوی شامل روش‌ها، ابزارها، الگوریتم‌ها یا مدل‌های مختلف است. جهت انجام عملیات داده‌کاوی نیاز هست که داده‌ها به شکل ساختاریافته باشند. این بدان معنی است که داده‌ها را می‌توان به‌صورت جدول یا یک پایگاه داده رابطه‌ای نشان داد. داده‌ها به شکل مجموعه‌ای از مثال‌ها (نمونه‌ها) با مقادیر خاص ویژگی‌های آن‌ها (ویژگی‌ها، متغیرها، زمینه‌ها) توصیف می‌شود. ویژگی‌ها می‌تواند انواع مختلفی داشته باشد که عبارتند از(دانگ و لیو[2]، ۲۰۰۴؛ لیو و متودا[3]، ۲۰۱۲):

- طبقه‌ای (اسمی)[4] - دامنه مجموعه‌ای گسسته از مقادیر است که در آن ترتیب معنی ندارد؛
- دو دویی[5] - نوع خاصی از ویژگی‌های طبقه‌ای که فقط با دو مقدار ممکن است؛
- ترتیبی[6] - دامنه یک مجموعه گسسته از مقادیر است که می‌تواند دارای یک ترتیب خاص باشد؛
- عددی[7] - مقدار یک عدد که این عدد صحیح یا پیوسته است.

یک واحد از متن می‌تواند یک جمله، چند جمله باهم ترکیب‌شده در یک پاراگراف یا متن‌های بسیار طولانی‌تری مانند صفحات وب، ایمیل، مقاله یا کتاب باشد. گاهی اوقات، یک متن می‌تواند فقط چند کلمه باشد که جمله معتبری نیست که کاملاً معمول است، به‌عنوان‌مثال، برای پست‌های کوتاه در شبکه‌های اجتماعی، برخی قوانین (نحو) ترکیب می‌شوند؛ بنابراین برای اینکه بتوان از روش‌های داده‌کاوی در متن استفاده کرد، باید آن‌ها را به نمایشی ساختاری تبدیل کرد.

[1] Witten and Frank
[2] Dong and Liu
[3] Liu H, Motoda
[4] categorical (nominal)
[5] binary
[6] ordinal
[7] numerical

به‌طورکلی می‌توان گفت که داده‌کاوی با داده‌های ساختاریافته و نرمال‌شده سروکار دارد و با پایگاه داده‌های رابطه‌ای کار می‌کند؛ اما در عوض متن‌کاوی با داده‌های ساختارنیافته یا نیمه‌ساختاریافته یعنی متون موجود در مقالات، اسناد و غیره سروکار دارد. علاوه بر این دسترسی کم به ساختار در متون دلیلی دیگر بر این است که متن‌کاوی کاری بسیار مشکل است. مفاهیم موجود در متون معمولاً بسیار انتزاعی بوده و به‌سختی می‌توان آن‌ها را مدل‌سازی موضوعی کرد. همچنین وقوع کلمات مترادف (کلمات متفاوت ازنظر نوشتاری اما هم‌معنی) یا کلمات با تلفظ یکسان (کلمات با تلفظ یکسان اما معنی متفاوت) پیدا کردن رابطه منطقی بین بخش‌های مختلف متن را مشکل می‌کند (اسپیناکیس و پریسترا[1]، ۲۰۰۴).

فرق داده‌کاوی با متن‌کاوی این است که در متن‌کاوی الگوها از متن زبان طبیعی استخراج می‌شوند درحالی‌که در داده‌کاوی الگوها از پایگاه‌های داده ساخت‌یافته به دست می‌آید. متن‌کاوی، متصل کردن اطلاعات استخراج‌شده به یکدیگر برای تشکیل حقایق یا فرضیه‌های جدید است تا پس‌ازآن به کمک روش‌های متعارف آزمایش، بررسی بیشتری شوند (لیو و متودا، ۲۰۱۲). به‌عبارت‌دیگر تفاوت میان داده‌کاوی و متن‌کاوی در این است که داده‌کاوی، اطلاعات را گردآوری و فهرست‌نویسی کرده، سپس اقدام به تولید دانش از بین حجم عظیمی از داده‌ها می‌کند اما متن‌کاوی، حوزه‌ای نو و میان‌رشته‌ای است که از رشته‌های بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی مشتق شده است و عمدتاً بر مستندات متنی تکیه دارد (رمضانی و همکاران، ۱۳۹۳).

۲-۲-۲-۲. تاریخچه متن‌کاوی

آنچه امروزه به‌طور عام متن‌کاوی نامیده می‌شود، مجموعه‌ای متشکل از علوم مختلف از قبیل زبان‌شناسی، آمار، کامپیوتر، مدیریت، هوش مصنوعی و دیگر حوزه‌ها ا .در اصل باید ریشه اصلی روش‌های متن‌کاوی را در تلاش‌ها و اقدامات صورت گرفته برای استفاده از فنون کمّی و آماری در علم زبان‌شناسی جست که به این مجموعه اقدامات،

[1] Spinakis and Peristera

زبان‌شناسی کمی[1] می‌گویند. تاریخچه استفاده از زبان‌شناسی کمی به حداقل قرن نوزدهم میلادی بازمی‌گردد. اما فعالیت کلاسیک تئوری گونهٔ زیف ۱۹۴۹ به‌عنوان یکی از مهم‌ترین پیش‌گامان عرصهٔ تحلیل کمی زبان‌شناسی، شناخته می‌شود (گلینسون[2] و همکاران، ۲۰۰۵). از دههٔ هفتاد میلادی تاکنون، افزایش چشمگیری در میزان علاقه به این حوزه از علوم اطلاعات مشاهده‌شده است. پژوهش ویلی یکی از نخستین کاربردهای این فنون در مطالعه منابع و ادبیات علمی، محسوب می‌شود (جانسنز[3] و همکاران، ۲۰۰۶).

نخستین بار لان[4] در سال ۱۹۵۸ میلادی، مفهوم متن‌کاوی را در مقاله خود مطرح نمود. سپس در سال ۱۹۶۱ نیز دویله در مقاله‌ای به متن‌کاوی و روش‌های مرتبط با آن اشاره کرد و بیان داشت که «طبقه‌بندی و سازمان‌دهی اطلاعات» می‌تواند از تجزیه‌وتحلیل تکرار و توزیع کلمات به‌کاررفته در آن صورت پذیرد. اگر این دو مورد را به‌عنوان نخستین موارد مطرح‌شدن متن‌کاوی در نظر بگیریم، می‌توان گفت که متن‌کاوی ممکن است مفهومی جدید باشد، اما رویای استخراج خودکار اطلاعات از متون، هم‌زمان با پیدایش کامپیوتر به‌وجود آمده است. سوانسون[5] در سال ۱۹۹۱ در مقاله خود بیان کرد که متون علمی باید به‌عنوان پدیده‌های ارزشمند طبیعی، کشف، اصلاح و تحلیل شوند و به‌این‌ترتیب نظر دانشمندان به استفاده از اطلاعات با تحلیل هوشمندانه را جلب کرد.

سوانسون با طراحی نرم‌افزاری نخستین گام در جهت عملی‌سازی متن‌کاوی را طی کرد. او از این نرم‌افزار در استخراج اطلاعات مفید از متون پزشکی استفاده نمود. این نرم‌افزار ارو اسمیس نام دارد. این نرم‌افزار که به‌صورت تخصصی در ارتباط با پردازش متون پزشکی تهیه شده قادر است پس از دریافت متون، کلمات کلیدی و متون مرتبط با یکدیگر را مشخص کند. سانسون هیچ‌گاه از اصطلاح متن‌کاوی برای این نرم‌افزار استفاده نکرد، اما به نظر می‌رسد که این نرم‌افزار نخستین نرم‌افزار متن‌کاوی بوده است. ازاین‌رو می‌توان او را پدر متن‌کاوی مدرن نامید.

---

[1] Quantitative linguistics
[2] Glenisson
[3] Janssens
[4] Luhn
[5] Swanson

لیندزی[1] و گوردون[2] ۱۹۹۹ کارهای سوانسون را بدون آنکه نام متن‌کاوی بر آن نهند، ادامه دادند. نرم‌افزار آن‌ها دو واژه را به‌صورت هم‌زمان در میان متون جستجو می‌کرد و نتیجه را در فهرستی قرار می‌داد تا کاربران آن‌ها را به‌عنوان مقاله‌های تکمیلی موردمطالعه قرار دهند. لیندزی و گوردون در همان سال روش TF-IDF[3] را برای رتبه‌بندی متون و واژه‌ها به این نرم‌افزار افزودند.

هیرست[4] (۱۹۹۷) نیز متن‌کاوی را با دسترسی به اطلاعات(بازیابی سنتی اطلاعات) متمایز ساخته است. بازیابی سنتی اطلاعات، بیشتر روی بازیابی متون مرتبط باهم تأکید می‌کند. متونی که با نیازهای اطلاعاتی کاربر مرتبط است. بر اساس تعریف هیرست داده‌کاوی-که متن‌کاوی نیز نوعی از داده‌کاوی به شمار می‌آید- تنها با اطلاعات و بازیابی آن سروکار ندارد، بلکه تلاش می‌کند تا اطلاعات جدیدی را از میان داده‌ها کشف نماید که پیش‌ازاین حتی برای ایجادکننده داده‌ها هم مشخص نبود. او معتقد بود که داده‌کاوی و متن‌کاوی مبتنی برشانس هست، درحالی‌که بازیابی اطلاعات مبتنی بر هدف است؛ بنابراین کارهایی نظیر جستجوی واژه‌ها برای پاسخ به پرسش‌ها، متن‌کاوی محسوب نمی‌شود. او بر این باور بود که بازیابی و دستیابی به اطلاعات می‌تواند به‌عنوان کاری تکمیلی و پشتیبان برای متن‌کاوی به شمار رود.

نخستین محصول نرم‌افزاری[5] در زمینه متن‌کاوی در سال ۱۹۹۹ توسط شرکت آی بی ام[6] به بازار عرضه گردید. نرم‌افزار مذکور شامل مجموعه ابزارهایی است که به استخراج اطلاعات از متن پرداخته و موجب غنای متن می‌شود. اطلاعاتی که از محتوای متن استخراج می‌شود می‌تواند ویژگی‌های متن نظیر زبان متن، اسامی افراد، تاریخ‌ها، مقادیر پول و ... باشد. استخراج این ویژگی‌ها به‌صورت خودکار بوده و بر اساس واژه‌نامه از پیش تعریف‌شده‌ای انجام نمی‌شود. این نرم‌افزار قادر است تا روی یک متن یا مجموعه‌ای از متون پردازش‌های لازم را انجام دهد. در این نرم‌افزار که شمارش واژه‌های مورد استفاده در متن،

[1] Lindsay
[2] Gordon
[3] Term Frequency–Inverse Document Frequency
[4] Hearst
[5] Intelligent Miner for Text
[6] IBM

اساس پردازش‌های بعدی آن است قسمتی برای تشخیص اصطلاح‌ها دارد و متون را خوشه‌بندی و دسته‌بندی می‌کند (در[1] و همکاران، ۱۹۹۹).

با توسعه مفهوم متن‌کاوی، مفاهیم دیگری نیز پابه‌پای آن رشد کرد: بازیابی اطلاعات از مجموعه متون، بازیابی اطلاعات از یک متن، کشف دانش از بانک‌های اطلاعاتی، مدیریت دانش در سازمان‌ها و مصورسازی داده‌ها و اطلاعات، این مفاهیم توسط کوستوف[2]، اوراد[3] و لوزیویچ[4] در سال ۲۰۰۰ در چند مقاله منتشر شد تا میان این مفاهیم تمایز قائل شوند. در سال ۲۰۰۱میلادی کوستوف و دمارکو[5] متن‌کاوی را با «استخراج اطلاعات از متون فنی» تعریف کردند. بر اساس این تعریف، متن‌کاوی شامل سه بخش است: بازیابی اطلاعات، پردازش اطلاعات و یکپارچگی اطلاعات. پردازش اطلاعات به استخراج الگوهای موجود در متون بازیابی شده اطلاق می‌شود و یکپارچگی اطلاعات ترکیبی از اطلاعات پردازش‌شده توسط کامپیوتر است که انسان اطلاعات را بازیابی نموده و مطالعه می‌کند. این مرحله مفهوم سیستم انسان- ماشین را تداعی می‌کند.

۳-۲-۲-۲. روش‌شناسی متن‌کاوی

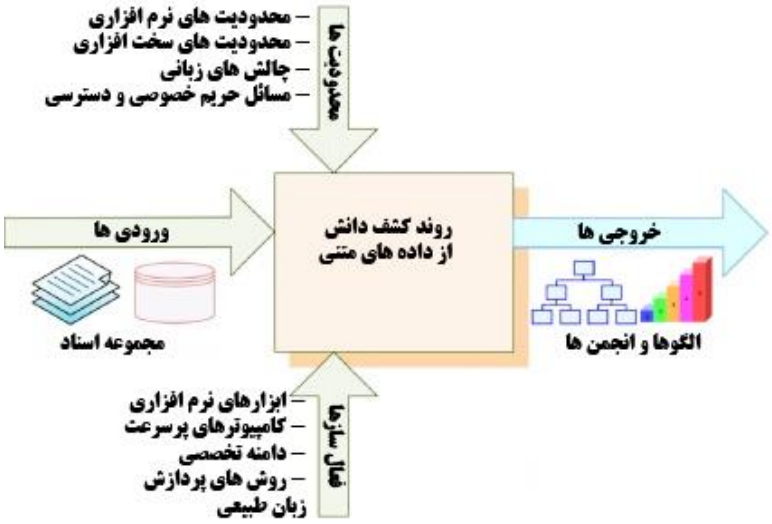
متن‌کاوی فرآیند نیمه‌خودکاری است که برای استخراج الگوها و کشف دانش از روی انبوهی از منابع داده‌ای غیر ساختاریافته مورد استفاده قرار می‌گیرد (دلن[6]، ۲۰۱۴). متن‌کاوی با داده‌کاوی رابطه تنگاتنگی دارند و هدف کُلی و برخی فرآیندهای عملیاتی آن‌ها یکسان است. همچنین در متن‌کاوی ورودی فرآیندها مجموعه‌ای از فایل‌های متنی غیر ساختاریافته بوده که از آن جمله می‌توان به فایل‌های Word، PDF، گزیده‌ای از متون و فایل‌های XML اشاره نمود. مزایای متن‌کاوی بیشتر در قلمروهایی است که انبوهی از داده‌های متنی قرار دارد. ازجمله مهم‌ترین این قلمروها می‌توان به

[1] Dörre
[2] Kostoff
[3] Oard
[4] Losiewicz
[5] DeMarco
[6] Delen

بخش مرور پیشینهٔ مقالات مختلف، بخش مالی (گزارش‌های فصلی، تفسیرهای رسانه‌ای)، خدمات بهداشتی (گزارش مرخصی بیمار، یادداشت‌های پزشک)، حقوق و مسایل قانونی (دستورات دادگاه)، زیست‌شناسی (فعل و انفعالات مولکولی)، فناوری (فایل‌های پروانه ثبت اختراع) و بازاریابی (نظرات مشتریان) اشاره نمود (کیم و دلن، ۲۰۱۸).

۴-۲-۲-۲. فرآیند متن‌کاوی

مطالعات متن‌کاوی با فرآیند اصولی و کارآمد و با بهره‌گیری از بهترین رویه‌های کاربردی بنانهاده می‌شود. هر پروژه‌ی متن‌کاوی به فرآیند استاندارد و موردپذیرش مجریان این قلمرو، همچون فرآیند استاندارد صنعتی Cross برای داده‌کاوی (CRISP-DM) نیاز دارد. در سطوح بالاتر فرآیند متن‌کاوی از طریق نمودار نمایش داده می‌شود نمونه چنین نموداری به همراه اجزاء تشکیل‌دهنده شامل ورودی‌ها، خروجی‌ها، کنترل‌گرها و سازوکارها (توانمندسازی‌ها) با پیکان‌های جهت‌دار در نمودار ۲-۲ نمایش داده‌شده‌اند (کیم و دلن، ۲۰۱۸).



نمودار ۲-۲- فرآیند متن‌کاوی در قالب یک مدل مفهومی کلی (کیم و دلن، ۲۰۱۸)

نمودار در شرایط مختلف به سه مرحله تقسیم می‌شود. هر یک از این مراحل دارای ورودی‌های خاصی بوده که برای تولید خروجی‌های اختصاصی بکار گرفته می‌شود. درصورتی‌که به هر دلیلی خروجی بخشی از این فرآیندها با انتظارات سازگاری نداشته باشد، اعمال بازخورد (بازگشت رو به عقب) نسبت به وظیفه‌ی در دست اجرا ضروری خواهد بود. در نمودار ۲-۲ نمای گرافیکی از این وضعیت نمایش داده‌شده است.

مرحلهٔ نخست. تعیین بخش‌های منتخب از متن: هدف اصلی فعالیت نخست گردآوری تمامی اسناد مرتبط با قلمروی در دست مطالعه است. در ادامه‌ی کار اسناد گردآوری‌شده به قالب خاصی تبدیل و در همان شرایط بایگانی‌شده تا همگی در ساختاری واحد تحت پردازش‌های رایانه‌ای متعاقب قرار بگیرند (برای مثال فایل‌های متنی در فرمت ASCII)؛

مرحله دوم. پیش‌پردازش داده‌ها (تشکیل ماتریس واژه / تفکیک برحسب سند): پس از نهایی سازی وضعیت متون، ماتریس واژه‌های تفکیک‌شده برحسب سند (TDM) بر اساس اطلاعات در دسترس تشکیل می‌شود. در این ماتریس سطرها نشان‌دهنده اسناد و ستون‌ها بیانگر واژه‌ها (کلیدواژه‌های مدنظر، هدف، جستجو شده) است. روابط برقرارشده بین واژه‌ها و اسناد به کمک شاخص‌ها نمایش داده می‌شوند (برای مثال شاخص‌های نسبی که ساده‌ترین آن‌ها تعداد تکرارپذیری هر واژه در اسناد مربوطه است)؛

مرحله سوم. استخراج دانش: هنگامی‌که به ماتریس TDM با ساختاردهی بهینه دسترسی باشد، به استخراج الگوها مبادرت ورزیده و آن‌ها به‌عنوان خوشه‌ها تعریف می‌شود. خوشه‌بندی یک فرآیند غیر نظارتی بوده که طی آن اشیاء یا رخدادها در گروه‌بندی‌های نرمال (معنادار) دسته‌بندی می‌شوند. فرآیند غیر نظارتی فرآیندی است که از هیچ‌گونه الگو یا دانش قبلی برای هدایت فرآیند خوشه‌بندی سود نخواهد بُرد. در فرآیند خوشه‌بندی غیر نظارتی، مجموعه اشیاء/ مؤلفه‌های فاقد برچسب‌گذاری (اسناد، دیدگاه‌های مشتری، صفحات وب) بدون هرگونه دانش قبلی به خوشه‌های معنادار منتقل می‌شوند. فرضیهٔ بنیادی بیانگر این نکته بوده که اسناد مرتبط کاملاً شبیه یکدیگر هستند، این در حالی است که اسناد غیر مرتبط از کمترین شباهت نسبت به یکدیگر سود می‌برند. درصورتی‌که این فرضیه معتبر باشد، آنگاه خوشه‌بندی اسناد بر مبنای میزان مشابهت محتوی به ارتقای کیفیت جستجو می‌انجامد.

یافتن تعداد بهینه‌ای از خوشه‌ها به‌هیچ‌عنوان وظیفه‌ای ساده نخواهد بود. درواقع هیچ‌گونه فرمول ریاضیاتی برای انجام این عمل پیش‌بینی‌نشده است (یک الگوریتم با ساختار بسته). تعیین تعداد بهینه‌ای از خوشه‌ها همچنان در قلمروی فرآیندهای شهودی-آزمایشی قرارگرفته و طی آن تعداد خوشه‌ها به‌تدریج از تعداد کم تا تعداد بیشتر افزایش‌یافته (یا برعکس) تا زمانی که تعداد خوشه‌های حاصله به شیوه‌ای بهینه بیانگر مجموعه داده‌ی چندبعدی در دست ارزیابی باشند (کیم و دلن، ۲۰۱۸).

۳-۲-۲. خوشه‌بندی

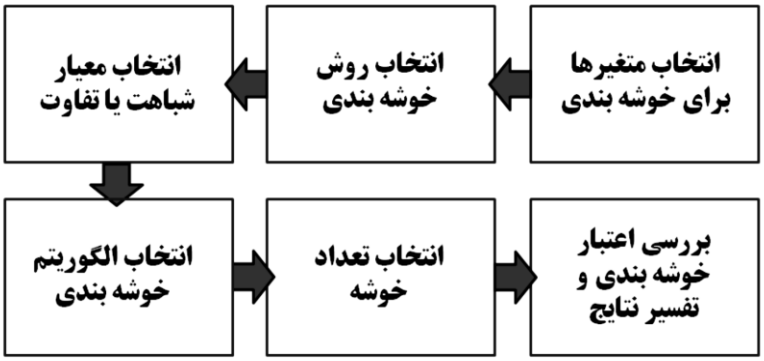
در سال‌های اخیر، به دلیل پیشرفت‌هایی که در داده‌کاوی، قدرت رایانه‌ها و بسته‌های نرم‌افزاری آماری حاوی الگوریتم‌های تحلیل خوشه، صورت گرفته است، تحلیل خوشه به موضوعی مهم مبدل شده است. اغلب نیاز است تا مجموعه‌ای از اسناد به گروه‌ها یا خوشه‌هایی همگن تقسیم‌بندی شوند. اگر این مجموعه حاوی تعداد اندکی مستند باشد، این کار می‌تواند به‌صورت دستی انجام گیرد. ولی اگر با حجم زیادی از اسناد سروکار داشته باشیم، انجام دستی این فرآیند، زمان‌بر و غیر مؤثر، خواهد بود. داده و الگو یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه‌شده است. قابلیت آن در ورود به فضای داده و تشخیص ساختار آن‌ها، خوشه‌بندی را یکی از ایده‌آل‌ترین سازوکارها برای کار با دنیای عظیم داده‌ها کرده است. ایده آن نخستین بار در دهه ۱۹۳۰ ارائه شد و امروزه با پیشرفت‌ها و جهش‌های عظیمی که پدید آمده، خوشه‌بندی در کاربردها و جنبه‌های مختلفی حضور یافته است. هدف نهایی خوشه‌بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم‌بندی داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند. البته کیفیت نتایج خوشه‌بندی به روش اندازه‌گیری شباهت و توانایی و قدرت الگوریتم در کشف الگوهای مخفی میان داده‌ها بستگی دارد (بورگلاند[1]، ۲۰۱۳).

[1] Borglund

خوشه‌بندی یکی از کاربردهای متن‌کاوی است. خوشه‌بندی به فرآیند تقسیم مجموعه‌ای از داده‌ها (یا اشیاء) به زیر کلاس‌هایی با مفهوم خوشه اطلاق می‌شود. به‌این‌ترتیب یک خوشه شامل یک سری داده‌های مشابه است که همانند یک گروه واحد، رفتار می‌کنند (بورگلاند، ۲۰۱۳؛ هانگ[1]، ۲۰۰۸). تکنیک خوشه‌بندی تکنیکی قابل‌اعتماد است که به‌طورکلی برای تحلیل کلان داده‌ها مورداستفاده قرار می‌گیرد. ثابت‌شده است که خوشه‌بندی متن یکی از مؤثرترین ابزارهای مورداستفاده برای تحلیل عنوان‌ها است (کلیفتون[2] و همکاران، ۲۰۰۴). هر مجموعه که موجودیت نامیده می‌شود، با یک خوشه مشخص می‌شود که مربوط به یکی از عنوان‌ها در مجموعه نوشته‌هاست (سالوم و همکاران، ۲۰۱۸).

خوشه‌بندی به‌عنوان یکی از روش‌های داده‌کاوی توصیفی، تکنیکی برای گروه‌بندی مشاهدات به K گروه (خوشه) مختلف است؛ به‌طوری‌که داده‌های هر خوشه بالاترین درجه شباهت را داشته باشند و داده‌های متعلق به خوشه‌های مختلف از حداکثر درجه عدم شباهت برخوردار باشند (کافمن و راسو[3]، ۲۰۰۹). خوشه‌بندی یکی از تکنیک‌هایی است که به‌طور گسترده به‌منظور متن‌کاوی، شناسایی الگو، تحلیل صفحات وب و تحلیل بازار مورد استفاده قرار می‌گیرد. در تعریفی دیگر، خوشه‌بندی برای جدا کردن یک جمعیت غیر همگن، به تعدادی از زیرگروه‌های همگن، بدون طبقه‌های از پیش تعریف‌شده، استفاده می‌شود. گروه‌ها با توجه به مشابهت مفاهیم و موجودیت‌های درون خوشه‌ها و بر اساس متغیرهایی مشخص و محتوای تحلیل خوشه‌بندی می‌شوند (تراپی[4] و همکاران، ۲۰۱۱). مویی[5] و سارستد[6] ۲۰۱۱ مراحل شش‌گانه‌ای برای انجام فرآیند خوشه‌بندی معرفی کرده‌اند که در نمودار ۲-۳. نشان داده‌شده است.

[1] Huang
[2] Clifton
[3] Kaufman and Rousseeuw
[4] Trappey
[5] Mooi
[6] Sarstedt



نمودار ۳-۲. مراحل شش‌گانه خوشه‌بندی (مویی و سارستد، ۲۰۱۱)

۱-۳-۲-۲. مراحل فرآیند خوشه‌بندی

- **انتخاب متغیرها برای خوشه‌بندی.** الگوریتم‌های خوشه‌بندی نتایج را فارغ از اینکه داده‌ها پیش‌فرض‌های لازم را برای خوشه‌بندی دارا باشند در اختیار کاربر قرار می‌دهد؛ بنابراین انتخاب متغیرهای صحیح مرتبط برای کسب نتایج معتبر بسیار حائز اهمیت است.
- **انتخاب روش خوشه‌بندی**. در این مرحله به طور معمول بهینه کردن مشخصه‌ای مانند به حداقل رساندن واریانس داخل خوشه‌ها یا حداکثر کردن فاصله بین خوشه‌ها انجام می‌شود.
- **انتخاب معیار شباهت یا تفاوت**. گام بعد برای انجام فرآیند خوشه‌بندی، انتخاب معیار تفاوت یا تشابه است تا بتوان فاصله دو مشاهده را به‌صورت یک مقدار عددی مشخص نمود (کافمن و راسو[1]، ۲۰۰۹). معیارهای متفاوتی برای خوشه‌بندی مجموعه داده‌ها وجود دارد که به ماهیت داده‌ها (کمی، اسمی و غیره) وابسته است (اوریت[2] و همکاران، ۲۰۱۱).

[1] kaufman & rasoi
[2] Everitt

- انتخاب الگوریتم خوشه‌بندی. بسته به نوع خوشه‌بندی (سلسله مراتبی، تقسیمی و غیره) از الگوریتم‌های مختلفی همچون الگوریتم پیوند، پیوند کامل، پیوند متوسط، الگوریتم مرکز، الگوریتم K-Means و غیره استفاده می‌شود.
- **انتخاب تعداد خوشه**. معیارهای اعتبار مختلفی برای ارزیابی خوشه‌بندی و انتخاب تعداد خوشه بر اساس محاسبات آماری وجود دارد. این معیارها بسته به روش خوشه‌بندی متفاوت است.
- **بررسی اعتبار خوشه‌بندی و تفسیر نتایج**. بررسی اعتبار خوشه‌بندی برای جلوگیری از ایجاد نتایج نامعتبر و خوشه‌های غیرحقیقی لازم است. توصیه می‌شود که پژوهشگران الگوریتم‌های مختلف خوشه‌بندی را برای داده‌های خود انجام داده و نتایج حاصل را بر اساس معیارهای اعتبار مقایسه نمایند تا با اطمینان بیشتری خوشه‌ها را انتخاب کنند. همچنین توصیه می‌شود تا خوشه‌بندی با تعداد خوشه‌های مختلف صورت گرفته و نتایج مقایسه شوند. استفاده از نظرات خبرگان حوزه موردنظر نیز در بررسی اعتبار نتایج خوشه‌بندی مفید خواهد بود (اوریت[1] و همکاران، ۲۰۱۱).

۴-۲-۲. مدل‌سازی موضوعی

به منظور برخورداری از روشی بهتر در جهت مدیریت بهینه اسناد الکترونیکی، استفاده از روش‌ها یا ابزارهایی که به‌طور خودکار عمل سازمان‌دهی، جستجو، نمایه‌سازی و مرور مجموعه‌های عظیم را انجام می‌دهد، ضروری است. براساس یافته‌های پژوهش‌های کنونی در مورد یادگیری ماشین و آمارها، تکنیک‌های جدیدی برای یافتن الگوهای واژه‌ها در مجموعه اسناد با استفاده از مدل‌های احتمالی سلسله مراتبی، توسعه‌یافته‌اند. این مدل‌ها، «مدل‌های موضوعی» نامیده می‌شود. کشف الگوها به‌طور معمول موضوعات اساسی به‌منظور شکل‌دادن اسنادی را منعکس کرده و به‌آسانی به سایر انواع داده‌ها تعمیم داده می‌شود. مدل‌های موضوعی به‌جای تحلیل واژه‌ها، تصاویر، داده‌های بیولوژیکی و اطلاعات و داده‌های نظرسنجی را تحلیل می‌کند(دیوید و همکاران، ۲۰۰۶).

اهمیت اصلی مدل‌سازی موضوعی، کشف الگوهای کاربرد واژه‌ها و چگونگی ارتباط با اسنادی است که الگوهای مشابه را ارائه می‌دهند (استیورز[1] و همکاران، ۲۰۰۷). در مدل‌سازی موضوعی هر سند موجود در یک مجموعه توسط هیستوگرامی که دربرگیرنده میزان وقوع واژه‌ها است، نمایش داده می‌شود. هیستوگرام با توزیع روی تعداد موضوع‌های معینی مدل‌سازی شده است که هر یک از آن‌ها توزیعی روی کلمات موجود در واژه‌نامه‌اند. با درک توزیع‌ها، می‌توان از هر سند، هیستوگرامی را ترسیم نمود. مدل‌های موضوعی مختلف ازجمله تحلیل معنایی پنهان[2]، تحلیل معنایی پنهانی احتمالاتی[3]، تخصیص پنهان دیریکله، مدل موضوعی هم‌بسته[4] طبقه‌بندی دقیقی در قلمرو کشف مدل‌سازی موضوعی ارائه داده‌اند (هافمن[5]، ۲۰۰۱). باگذشت زمان، موضوعات موجود در مجموعه اسناد رشد داشته و باید گفت مدل‌سازی موضوعی بدون در نظر گرفتن متغیر زمان باعث ابهام در کشف موضوع می‌شود. مدل‌سازی موضوعی با لحاظ نمودن متغیر زمان، مدل‌سازی تکاملی موضوعی نامیده می‌شود. این مدل‌سازی اطلاعات پنهان و مهم در مجموعه اسناد را افشاء نموده و امکان شناسایی موضوعات را باتوجه به زمان و بررسی تکاملی‌شان در طول زمان ارائه می‌دهد. قلمروهای مختلفی می‌توانند از مدل‌های تکاملی موضوعی استفاده کنند.

۱-۴-۲-۲. روش‌های مدل‌سازی موضوعی

در این بخش، در خصوص برخی از روش‌های مدل‌سازی موضوعی که با کلمات، اسناد و موضوعات سروکار دارند، بحث می‌شوند. به‌علاوه، ایده کلی این روش‌ها و مثال‌هایی در صورت لزوم ارائه شده است. این روش‌ها، کاربردهای زیادی دارند در اینجا چگونگی کاربرد هر روش به‌طور مختصر بیان می‌گردد.

---

[1] Steyvers
[2] Latent Semantic Analysis (LSA)
[3] Probabilistic Latent Semantic Analysis (PLSA)
[4] Correlated Topic Model (CTM)
[5] Hofmann

- **تحلیل معنایی پنهان**

تحلیل معنایی پنهان (LSA) روش یا تکنیک مربوط به پردازش زبان طبیعی است. هدف اصلی تحلیل معنایی پنهان، ایجاد بردار بازنمایی برای متون و نشانه‌گذاری محتوای معنایی است. با استفاده از بازنمایی برداری، شباهت میان متون محاسبه می‌شود. در گذشته، تحلیل معنایی پنهان، نمایه‌سازی معنایی پنهان[1] (LSI) نام داشت و برای بازیابی اطلاعات، ارائه شده بود. تحلیل معنایی پنهان بایستی ابعادی برای روشی مانند تطبیق کلمات کلیدی، تطبیق وزن کلمات کلیدی و بازنمایی بردار مربوط به رخداد کلمات در اسناد، داشته باشد (سیلوا و دی‌کاروالهو[2]، ۲۰۲۱). همچنین، تحلیل معنایی پنهان از تجزیه مقدارهای تکی[3] (SVD) برای مرتب‌سازی مجدد داده‌ها استفاده می‌نماید. تجزیه مقدارهای تکی روشی است که از یک ماتریس برای پیکربندی مجدد و محاسبه نقصان‌های فضای بردار، استفاده می‌شود. نقصان‌های فضای بردار از بالاترین تا کمترین اهمیت سازمان‌دهی می‌شوند. برای توصیف اساسی‌ترین مراحل تحلیل معنایی پنهان ابتدا، مجموعه عظیمی از متون مرتبط گردآوری‌شده و سپس توسط اسناد، تقسیم می‌شود. سپس، ماتریس هم‌رخدادی برای عبارات و اسناد می‌گردد. در ادامه، هر سلول انتخاب و محاسبه می‌شود. درنهایت باید افزود تجزیه مقدارهای تکی نقش مهمی برای محاسبه ابعاد و ایجاد ماتریس سه‌بعدی ایفاء می‌کند (القمدی و الفلقی[4]، ۲۰۱۵).

- **تحلیل معنایی پنهان احتمالی**

تحلیل معنایی پنهان احتمالی (PLSA) روشی است که پس از روش تحلیل معنایی پنهان و برای برطرف نمودن معایب تحلیل معنایی پنهان ارائه شد. تحلیل معنایی پنهان احتمالی روشی است که امکان نمایه‌سازی خودکار سند را بر اساس مدل کلاس پنهانی آماری افزایش می‌دهد و درصدد بهبود تحلیل معنایی پنهان به‌صورت احتمالاتی و با استفاده از مدل مولد

[1] Latent Semantic Indexing
[2] Silva & de Carvalho
[3] Singular Value Decomposition
[4] Alghamdi and Alfalqi

است. هدف اصلی تحلیل معنایی پنهان احتمالی، شناسایی و تمیز میان مفاهیم مختلف واژه‌های به‌کاررفته بدون استفاده از واژه‌نامه است. این روش دو نتیجه مهم دارد: نخست اینکه ابهام‌زدایی از واژه‌های دارای چند معنا را ممکن می‌سازد و دوم با گروه‌بندی واژه‌هایی که چارچوب مشترکی دارند، شباهت‌های معمول را نشان می‌دهد (هافمن، ۲۰۰۱).

تحلیل معنایی پنهان احتمالی مبتنی بر مدلی آماری است که در قالب مدلی ابعادی شناخته‌شده است. مدل ابعادی مدل متغیر پنهان برای داده‌های بارخداد مشترک است که دسته متغیرهای مشاهده نشده را با هر مشاهده، مرتبط می‌کند. روش تحلیل معنایی پنهان احتمالی برای بهبود تحلیل معنایی پنهان ارائه‌شده و بنابراین مشکلاتی که تحلیل معنایی پنهان نمی‌تواند حل کند را آدرس‌دهی می‌نماید. تحلیل معنایی پنهان احتمالی کاربردهای موفقی ازجمله بینایی رایانه‌ای[1] و سیستم‌های توصیه گر[2] در دنیای واقعی داشته است.

کاربردهای تحلیل معنایی پنهان احتمالی حوزه‌های مختلفی ازجمله بازیابی و فیلترسازی اطلاعات، پردازش زبان طبیعی و یادگیری ماشین از متن را شامل می‌شود. به‌ویژه، برخی از این کاربردها، طبقه‌بندی پردازش خودکار، دسته‌بندی، ردگیری موضوعی، بازیابی تصویر و توصیه خودکار پرسش هستند. در ادامه دو مورد از این کاربردها توضیح داده شده است.

- بازیابی تصویر. مدل تحلیل معنایی پنهان احتمالی ویژگی‌های بصری دارد که باعث می‌شود برای نمایش هر تصویر به‌عنوان مجموعه واژه‌های بصری از یک واژه‌نامه بصری به کار رود. وقوع واژه‌های بصری در یک تصویر در بردار رخداد مشترک محاسبه شده است. هر تصویر، بردارهای رخداد مشترکی دارد که به ساخت جدول رخداد مشترک برای به‌دست‌آوردن مدل تحلیل معنایی پنهان احتمالی کمک می‌کند. پس از شناخت مدل تحلیل معنایی پنهان احتمالی می‌توان آن‌را به همه تصاویر پایگاه داده اعمال نمود. سپس آرایش برداری برای نمایش هر تصویر به کار می‌رود و هر عنصر بردار درجه‌ای که هر تصویر، موضوع معینی را ترسیم می‌کند را نشان می‌دهد (رامبرگ[3] و همکاران، ۲۰۰۸)

[1] Computer vision
[2] recommender systems
[3] Romberg

- توصیه خودکار پرسش. یکی از کاربردهای مهمی که تحلیل معنایی پنهان احتمالی با آن سروکار دارد، وظیفه توصیه پرسش است. در این کاربرد، واژه مستقل از کاربر است. اگر کاربر معنای خاصی مدنظر داشته باشد و نیز زمانی که کاربر، پاسخ‌ها و معناهای نهفته مربوط به پرسش را دریافت می‌کند، می‌تواند توصیه‌هایی براساس شباهت‌های معناهای نهفته ارائه دهد. وو[1] و همکاران (۲۰۰۸) اظهار داشتند که تحلیل معنایی پنهان احتمالی به منظور مدل‌سازی پروفایل کاربران استفاده می‌شود. همچنین پرسش‌ها با حذف احتمالات موضوعات پنهان پشت واژه‌ها مدل‌سازی می‌شوند. دلیل این امر آن است که پروفایل کاربر توسط پرسش‌ها و پاسخ‌ها نمایش داده شده است. از این‌رو، فقط چگونگی مدل‌سازی درست پرسش مدنظر قرار می‌گیرد.

- **تخصیص پنهانی دیریکله**

بهبود روش مدل‌سازی ترکیبی علت اصلی بروز مدل تخصیص پنهانی دیریکله (LDA) بود که قابلیت تبادل واژه‌ها و اسناد را از روش قدیمی و توسط تحلیل معنایی پنهان احتمالی و تحلیل معنایی پنهان ثبت می‌کند. این امر همزمان با نظریه بازنمون کلاسیک[2] در سال ۱۹۹۰روی داد (بلی و همکاران، ۲۰۰۳).

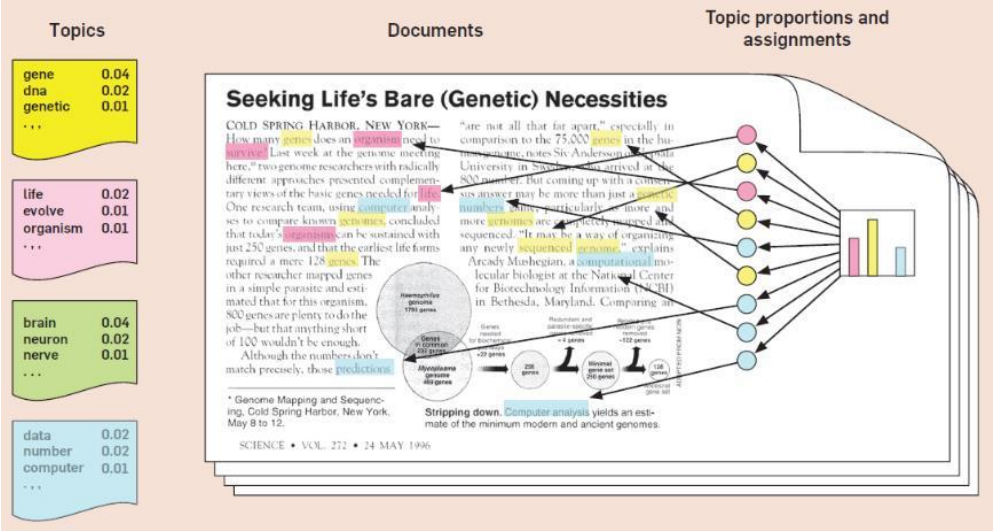
مجموعه اسناد الکترونیکی متعدد از جمله صفحات وبی، وبلاگ‌های علمی و مقالات اخباری چالش‌های جدیدی برای پژوهشگران در جامعه داده‌کاوی ایجاد کرده‌اند. ضرورت روش‌های خودکار بصری‌سازی، تحلیل و خلاصه‌سازی این مجموعه اسناد افزایش‌یافته است. اخیراً، مدل‌سازی موضوعی پنهان به‌عنوان تکنیکی بدون نظارت، محبوبیت زیادی برای کشف موضوع در مجموعه کلان داده‌های متنی کسب کرده است. این مدل، الگوریتمی برای متن‌کاوی است که براساس مدل‌های موضوعی آماری (بیزین) بوده و کاربرد بسیار گسترده‌ای داشته است. مدل تخصیص پنهانی دیریکله یک مدل مولد است که در صدد تقلید پردازش نوشتن است؛ بنابراین سعی می‌کند یک سند را براساس موضوع موردنظر، تولید کند. این مدل همچنین می‌تواند برای سایر انواع داده‌ها به کار رود. ده‌ها مدل مبتنی بر مدل تخصیص پنهانی دیریکله وجود دارد از جمله متن‌کاوی موقتی،

[1] Wu
[2] Classic representation theorem

تحلیل موضوع – نویسنده، مدل‌های موضوعی نظارتشده، خوشه‌بندی مشترک دیریکله و مدل تخصیص پنهانی دیریکله مبتنی بر بیوانفورماتیک (شن[1] و همکاران، ۲۰۰۸؛ وانگ و همکاران، ۲۰۱۶).

شکل ۱-۲ ایده نهفته در منطق الگوریتم تخصیص پنهان دیریکله را نشان داده است. این روش فرض می‌کند، تعدادی موضوع که هرکدامشان توزیعی از کلمات است، در کل مجموعه مستندات موجود است (سمت چپ شکل). حال می‌توان تصور نمود که هر مستند این‌گونه شکل می‌گیرد:

ابتدا توزیعی از موضوعات انتخاب می‌شود، سپس برای هر واژه یک تخصیص موضوعی (دایره‌های رنگی در شکل) گزینش شده و در نهایت نیز واژه موردنظر از موضوع مربوط انتخاب می‌گردد. شایان‌ذکر است که موضوعات و تخصیص‌های موضوعی به نمایش درآمده در این تصویر تنها به‌عنوان‌مثالی توضیح‌دهنده ارائه‌شده و از اعمال بر داده‌های واقعی به دست نیامده‌اند (بلی، ۲۰۱۲).



شکل۴-۲. ایده نهفته در الگوریتم تخصیص پنهان دیریکله (گریفت و استیورز، ۲۰۰۴)

[1] Shen

ایده اصلی این فرآیند این است که هر سند به‌صورت ترکیبی از موضوعات به صورت مدل درآمده است و هر موضوع یک توزیع احتمالی گسسته است که تعیین می کند چگونه احتمال حضور هر کلمه در موضوع موردنظر وجود دارد. این احتمالات موضوعی، بازنمود دقیقی از سند را ارائه می‌کنند (برگولز[1] و همکاران، ۲۰۰۸). در اینجا، یک «سند» در واقع «کیسه‌ای از کلمات»[2] است که ساختاری فراتر از آماره های کلمه و موضوع ندارد. از کاربردهای مدل مبتنی بر روش مدل تخصیص پنهانی دیریکله می‌توان موارد زیر را نام برد:

- **کشف نقش**[3]: تحلیل شبکه اجتماعی (SNA) بررسی مدل‌های ریاضی برای تعامل میان افراد، سازمان‌ها و گروه‌ها است. به دلیل بروز ارتباطاتی میان مهاجمان امنیتی یازدهم سپتامبر و مجموعه داده‌های عظیم انسانی در سرویس‌های وب محبوب از جمله فیسبوک و مای اسپیس[4]، علاقه به تحلیل شبکه‌های اجتماعی، افزایش‌یافته است. این امر منجر به مدل موضوع - گیرنده - نویسنده[5] (ART) برای تحلیل شبکه‌های اجتماعی شد. مدل تخصیص پنهانی دیریکله و مدل- موضوع- نویسنده را ترکیب می‌کند. ایده مدل- موضوع - گیرنده - نویسنده، درک توزیع‌های موضوعی براساس پیام‌های حساس به جهت که بین دریافت‌کنندگان و فرستندگان ارسال شده است، دارای اهمیت می‌باشد (مک کلوم[6] و همکاران، ۲۰۰۷).
- **موضوع احساسی**[7]: مدل جفت- لینک[8]– مدل تخصیص پنهانی دیریکله که بر مسئله مدل‌سازی اتصال متن و ارجاعات در قلمرو مدل‌سازی موضوعی متمرکز است. این براساس ایده مدل تخصیص پنهانی دیریکله و مدل‌های بلوک احتمالی عضویت ترکیبی[9] (MMSB) است و امکان مدل‌سازی دلخواه لینک را می‌دهد (بائو[10] و همکاران، ۲۰۰۹).

[1] Bergholz
[2] bag of words
[3] Role discovery
[4] My space
[5] Author-Recipient-Topic
[6] McCallum
[7] Emotion topic
[8] Pairwise-Link-LDA
[9] Mixed Membership Stochastic Block Models
[10] Bao

- **نمره‌دهی خودکار متن**[1]: مسئله نمره‌دهی خودکار متن همبستگی نزدیکی با طبقه‌بندی متن دارد و از دهه ۱۹۶۰ مورد بررسی قرار گرفته است. مدل تخصیص پنهان دیریکله در مقایسه با روش‌های کاهش ابعاد[2] برای نمره‌دهی خودکار متن، نشان داده که رویکردهای قابل اطمینانی برای وظایف مربوط به بازیابی اطلاعات وجود دارد (کاکونن[3] و همکاران، ۲۰۰۶).

- **مدل موضوعی هم‌بسته**

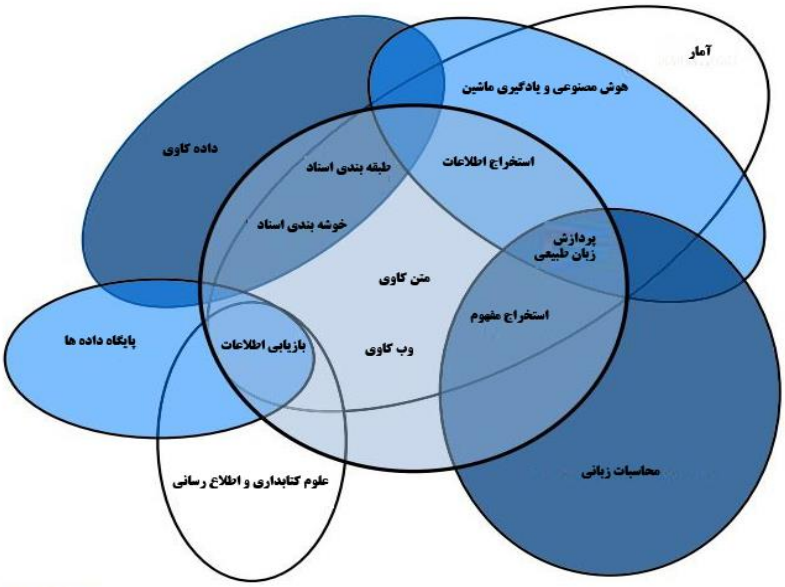
مدل موضوعی هم‌بسته[4] نوعی مدل آماری مورد استفاده در پردازش زبان طبیعی و یادگیری ماشین است. مدل موضوعی هم‌بسته برای کشف موضوعاتی که در گروه اسناد نشان داده‌اند، به‌کاررفته است. کلید مدل موضوعی هم‌بسته توزیع نرمال لجستیک است. مدل‌های موضوعی هم‌بسته به مدل تخصیص پنهانی دیریکله متکی است (القمدی و والفلقی، ۲۰۱۵).

۲-۴-۲-۲. ماهیت بین‌رشته‌ای متن‌کاوی

منظور از بین‌رشته‌ای بودن به این معنی است که با به‌کارگیری اصول و قواعد دانش‌های موجود (دانشی بیشتر از دو قلمرو علمی) و ترکیب آن‌ها با هم دانش جدید ایجاد می‌شود (وانگ و همکاران، ۲۰۲۰). متن‌کاوی زمینه‌ی بین‌رشته‌ای است که

[1] Automated essay scoring
[2] Dimension Reduction Methods
[3] Kakkonen
[4] Correlated Topic Model (CTM)

به‌طورکلی از شش قلمرو دانشی و اشتراک آنها بهره می‌گیرد. آمار، هوش مصنوعی و یادگیری ماشین، داده‌کاوی، محاسبات زبانی، پایگاه داده و علوم کتابداری و اطلاع‌رسانی رشته‌هایی هستند که متن‌کاوی از آن‌ها استفاده می‌کند. شکل ۵-۲ نمودار اشتراکات متن‌کاوی و سایر رشته‌ها را نشان می‌دهد (ماینر و همکاران، ۲۰۱۲).



شکل ۵-۲. نمودار ون اشتراکات متن‌کاوی و سایر قلمروهای علمی (ماینر و همکاران، ۲۰۱۲)

همان‌طور که در شکل ۵-۲ مشاهده می‌شود از اشتراک رشته‌های مذکور، هفت روش عملی برای رسیدن به اهداف متن‌کاوی ایجاد می‌شود که عبارتند از پردازش زبان طبیعی، طبقه‌بندی اسناد، خوشه‌بندی اسناد، بازیابی اطلاعات، وب کاوی و استخراج مفهوم هستند.

- طبقه‌بندی اسناد[1]: در این روش بهترین برچسب از مجموعه برچسب‌های از قبل مشخص شده برای مدارک بدون برچسب[2] مورد استفاده قرار می‌گیرد.

- خوشه‌بندی اسناد[3]: به فرآیند گروه‌بندی مستندات مشابه درون خوشه‌های مختلف خوشه‌بندی می‌گویند.

[1] Document Classification
[2] untagged documents
[3] Document Clustering

- بازیابی اطلاعات[1]: فرآیند نمایه‌سازی[2]، جستجو و بازیابی و بازیابی مستندات از بین مجموعه داده‌های عظیم متنی با استفاده از کلیدواژه‌ها را بازیابی اطلاعات گویند. مهم‌ترین کاربرد بازیابی اطلاعات در موتورهای جستجوی دیده می‌شود.

- وب کاوی[3]: فرآیند داده‌کاوی و متن‌کاوی بر روی محتوای صفحات وبی و لینک‌ها را وب کاوی می‌گویند. صفحات وب حالتی نیمه‌ساختاریافته دارند و از مجموعه‌ای از متون و لینک‌ها به صفحات دیگر تشکیل شده‌اند. لذا یک رویکرد متداول برای وب‌کاوی، بازنمایی صفحات وب در قالب گراف و تحلیل گراف‌های وبی است. امروزه زیرشاخه‌ای از وب‌کاوی برای تحلیل شبکه‌های اجتماعی مانند فیس‌بوک، توییتر و ... بسیار موردتوجه پژوهشگران قرار گرفته است.

- استخراج اطلاعات[4]: فرآیند شناسایی و استخراج موجودیت‌های مناسب و همچنین روابط بین آن‌ها از درون متن (غیر ساخت‌یافته) را استخراج اطلاعات می‌گویند. به‌عبارت‌دیگر، استخراج اطلاعات فرآیندی برای پردازش داده‌های غیر ساخت‌یافته (مثل متن، تصویر، صوت و ...) و نیمه‌ساختاریافته (مثل صفحات وب، XML، ...) و تبدیل آنها مجموعه داده ساختاریافته (از قبیل جداول پایگاه‌داده) است. استخراج اطلاعات به دو نوع باز (عمومی) و بسته (خاص و در حوزه‌ای مشخص) تقسیم می‌شود.

- پردازش زبان طبیعی[5]: هدف آن، پردازش سطح پایین و فهم (درک) زبان و بخصوص متن توسط کامپیوترها است. معمولاً معادل با اصلاح زبان‌شناسی رایانشی[6] بکار گرفته می‌شود، هرچند که اغلب زبان‌شناسان زبان‌شناسی رایانشی را کلی‌تر از پردازش زبان طبیعی می‌دانند.

[1] Information Retrieval-IR
[2] indexing
[3] Web Mining
[4] Information Extraction-IE
[5] Natural Language Processing-NLP
[6] Computational linguistics

- استخراج مفاهیم[1]: فرآیند گروه‌بندی واژه‌ها و عبارات درون گروه‌های متنی که دارای مشابهت معنایی هستند را استخراج مفاهیم گویند. معمولاً از تکنیک‌های آماری (مانند n-grams)، مدل‌سازی موضوعی[2] و خوشه‌بندی متون و واژه‌ها برای استخراج مفاهیم استفاده می‌شود.

۳-۲. مرور پیشینه‌های پژوهش

در دهه‌های گذشته، چالش اصلی در فرایند تصمیم‌گیری و انتخاب، کمبود اطلاعات بود؛ اما امروزه اشباع اطلاعاتی و انتخاب اطلاعات مناسب، چالش اصلی به شمار می‌آید. فضای تولیدات علمی به صورت مقاله، ثبت اختراع و اکتشاف، چاپ کتاب، انتشار مجلات علمی–تخصصی، چه از نظر اقتصادی و چه از جنبه های مدیریتی، رعایت اخلاق و امانتداری علمی، تسهیم منافع علمی، آموزش آداب و اخلاق علمی عدم ضابطه مندی را نشان می دهد. این عدم ضابطه مندی، بر آشفتگی در تولید علم دامن می زند (کینی، ۲۰۰۷). از طرفی از نظر عملیاتی، نویسندگان با شناسایی و طبقه بندی میدانی نشریات و انتساب ارزش و ارتباط موضوعی هر نشریه اقدام به ارسال و انتشار یافته های پژوهش خود می کنند. از اینرو این نکته که نشریات تا چه اندازه در اعلام حوزه های موضوعی که مقالات را میپذیرند شفاف عمل کرده و به آن پایبند بوده حائز اهمیت است. گاها موضوعات اعلامی از سوی نشریات بسیار کلی است و کاربر را در انتخاب نشریه مرتبط با دشواری مواجه می کند. از طرفی با توجه به محدودیت‌های زمانی و همچنین حجم اطلاعات، جستجو و تحلیل دستی منابع علمی عملا امکان‌پذیر نیست و آنچه حائز اهمیت است استفاده از روش‌های یادگیری ماشین مانند متن کاوی برای پردازش داده‌های بزرگ است. از جمله کاربردهای متن‌کاوی می‌توان به دسته‌بندی، خوشه‌بندی، خلاصه‌سازی و یافتن روابط میان مفاهیم در متون اشاره کرد. متن‌کاوی دارای دو روش یادگیری«با نظارت[3]» و «بدون نظارت[4]» است (الهیاری، ۲۰۱۷). در روش یادگیری با نظارت، هر یک از داده‌های آموزشی به دسته‌ای خاص که از ابتدا مشخص هستند نسبت داده می‌شود. بعبارتی در هنگام آموزش ناظری

[1] Concept Extraction
[2] topics modeling
[3] supervised
[4] unsupervised

وجود دارد که اطلاعاتی علاوه بر داده‌های آموزش در اختیار یادگیرنده قرار می‌دهد. در یادگیری بدون نظارت، یادگیرنده است که باید درون داده‌ها به دنبال ساختاری خاص بگرد؛ زیرا به‌جز داده‌های آموزشی، هیچ اطلاعاتی در اختیار یادگیرنده قرار ندارد (کرابر و همکارن[1]، ۲۰۱۹).

در پیوند با پژوهش حاضر طبقه بندی یا دسته بندی موضوعی مقالات نشریه ها به گونه ای مد نظر است که بتواند مفهوم را به درستی منتقل نماید. از اینرو از روشهای مدلسازی موضوعی بهره برده خواهد شد. مدلسازی موضوعی تکنیکی است که ساختار موضوع را در مجموعه ای از اسناد کشف و تفسیر مینماید (بلی[2]، ۲۰۱۲). به عبارت دیگر، در حالت کلی روشها و الگوریتمهایی هستند که متن را پردازش کرده و موضوعات مختلف موجود در آن را استخراج مینمایند.

االگوریتم‌ها برای تولید و ارائه پیشنهاد از عناصر متنی مقالات مانند عنوان، چکیده و واژه‌های کلیدی استفاده می‌کنند (آگاروال، ۲۰۱۶). که به شکل انبوهی از داده‌های ساخت‌نیافته در متون وجود دارند و بایستی به نحوی که که برای ماشین قابل استفاده و درک باشد استخراج شود (کوچوکتانچ[3] و همکاران، ۲۰۱۲). بنابراین میتوان گفت که در مدلسازی موضوعی هر سند با توجه به موضوعات موجود در آن تفسیر و سازماندهی میگردد .در این مدلها، هر متن یا سند به صورت توزیعی از موضوعات ارائه میگردد؛ در حالی که هر موضوع هم توزیعی روی واژگان میباشد (کروا و بانسل، ۲۰۲۰). در مدلسازی موضوعی موضوعات با تشخیص الگوهایی مثل کلمات و تکرارشان تشخیص داده می شوند. مدل فضای برداری (VSM) اولین مدل جبری ساده ای بود که مستقیما بر اساس ماتریس اصطلاح- سند برای استخراج اطلاعات معنایی ارائه گردید (سالتون و همکاران، ۱۹۷۵). یک مدل فضای برداری پایه میتواند متن را به کاراکترهای یونی گرم تا ان-گرم تقسیم کند که بر اساس روش کیف کلمات (BOW) میباشد. به عبارت دیگر، ترتیب دقیق اصطلاح در یک سند نادیده گرفته شده؛ اما تکرار وقوع هر اصطلاح به عنوان یک فاکتور مهم نگهداری میگردد (بری و همکاران، ۱۹۹۹). کاربرد مدل فضای برداری در حوزه هایی است که نیاز به محاسبه میزان شباهت میان کلمات و اصطلاحات موجود در اسناد دارند که از آن

[1] Kruber et al.
[2] Blei
[3] Küçüktunç

جمله میتوان به موتورهای جستجو، پردازش زبان طبیعی و ماشین اشاره کرد (گرب و همکاران، ۲۰۱۳).

برای مدلسازی سامانه بازیابی مانند موتورهای جستجو یا مجموعه دیجیتالی، همه کلمات در یک سند به یک اندازه مهم نیستند بنابراین به هر اصطلاح در سند یک وزن بر اساس تعداد وقوع اصطلاح در آن سند داده میشود. این وزن بهتر است بر اساس وقوع یک اصطلاح در سند و عدم تکرار در اسناد دیگر باشد .بنابراین در متون معمولا از وزندهی Tf_IDF استفاده میگردد (تورنی و پانتل، ۲۰۱۰).

الگوریتمهای مدلسازی موضوعی با استفاده از تکنیکهای مختلف تلاش میکنند تا موضوعات خوشه بندی شده را تحت عنوان یک موضوع ارائه دهند. مدل توزیعی مانند فضای برداری، آنالیز پنهان مفهوم ، آنالیز پنهان مفهومی احتمالی و تخصیص دیریکله پنهان میتواند معنای کلمات را از متن با استفاده از روشهای آماری استخراج کند (کرین و همکاران، ۲۰۱۲، زمانی و همکاران، ۱۳۹۳). به منظور انجام مدلسازی موضوعی در متن، نیاز به پردازش زبان میباشد. ابزارهای استاندارد پیش‌پردازش و نرمالسازی ایجاد شده برای متون زبان فارسی به صورت رایگان منتشر نشده یا دقت مناسب را ندارند (کامیابی و همکاران، ۱۳۹۷). به دلیل نزدیکی دبیره زبان فارسی با عربی، همواره در نگارش تعدادی از حرف‌ها مشکل نویسه‌های عربی معادل وجود دارد. از جملهٔ آنها می‌توان به حروف «ک»، «ی»، همزه و... اشاره کرد. علاوه بر این، وجود نویسهٔ «ـّ»، تشدید، تنوین و موارد مشابه (کامیابی و همکاران، ۱۳۹۷؛ سراجی[1]، ۲۰۱۰) ایست واژه‌ها[2] از جمله حروف اضافه، بسیاری از قیدها، حروف ربط و افعال از جمله اقدام‌های لازم قبل از شروع پردازش متن است که این پردازشها در هر سطح، نیازمند دانش، منابع و پیکره های مورد نیاز آن سطح و سطوح پایینتر است. در دسترس بودن منابع و دانش برای انجام تحقیق در حیطه ی پردازش زبان طبیعی از جمله چالشهای پردازش زبان طبیعی است.

در بحث متن‌کاوی و پردازش زبان طبیعی از شیوه‌ها و الگوریتم‌های متفاوتی به‌منظور وزن‌دهی به واژگان استفاده می‌شود. هر یک از این الگوریتم‌ها دارای نقاط قوت و ضعفی هستند که براساس روش‌شناسی پژوهش‌های گوناگون و نیز شرایط حاکم بر پژوهش‌ها برای وزن‌دهی به واژگان مورد استفاده قرار می‌گیرند. مرور پژوهش‌هایی که از روش TF-IDF برای وزن‌دهی

[1] Seraji
[2] Stop word

استفاده کرده‌اند، حاکی از آن است که این روش از نقاط قوت بسیار زیادی برخوردار است (راموس[1]، ۲۰۰۳؛ ایلیاس[2]، ۲۰۱۹؛ کیم و دلن، ۲۰۱۸). راموس (۲۰۰۳) یکی از مهم‌ترین مزایا و کاربرد TF-IDF را در حجم بالای داده‌ها می‌داند. به بیان دیگر، استفاده از TF-IDF در پژوهش‌هایی که دارای داده‌هایی با حجم بالا هستند، توصیه می‌شود.

پژوهشهای مرتبط با بررسی میزان تخصص‌گرایی یا تنوع موضوعی مقالات نشریات با استفاده از روشهای متن‌کاوی به صورت موردی انجام شده است. در پیوند با پژوهش حاضر به دلیل تفاوتهای ذاتی زبان فارسی نسبت به سایر زباتها، در این بخش به پژوهشهای مرتبط با زبان فارسی پرداخته شده است.

تحلیل محتوای موضوعی یا ترسیم نقشه موضوعی مقالات با استفاده از روشهای متن کاوی به صورت موردی برای برخی از نشریه های علمی انجام شده است. نتایج مطالعه افضلی پور و جمالی مهموئی در بررسی میزان تخصص گرایی مجله های حوزه علم اطلاعات و دانش شناسی فارسی طی سالهای ۱۳۸۷ الی ۱۳۹۱ نشان داد که ۹۳ درصد مقالات با موضوعهای مرتبط و ۷ درصد مقالات با موضوعای غیرمرتبط با علم اطلاعات در این نشریات منتشر شده اند. همچنین نتایج نشان داد در غالب نشریات تناسبی بین موضوعی و نام نشریه وجود ندارد. در این پژوهش از طبقه بندی موضوعی لیزا استفاده شده است (۱۳۹۳). . ناغانی و عابسی (۱۳۹۵) ازروش متن کاوی به منظور آنالیز محتویات مقالات علمی در زمینه مهندسی صنایع استفاده و تکنیکهایی را ارائه دادند.

شکوهیان و همکاران (۱۳۹۸) به ارائه مدل دسته بندی موضوعی تولیدات علمی حوزه سلامت با استفاده از روشهای متن کاوی پرداختند. در این مطالعه از مدلی ترکیبی برای دسته بندی موضوعی تولیدات علمی استفاده شده است. داده ها مربوط به سال ۲۰۰۹ الی ۲۰۱۹ مستخرج از پایگاه پابمد می باشد. نتایج نشان داد ترکیب دسته بندی و خوشه بندی می تواند دقت دسته بندی متون سلامت را افزایش دهد.

[1] Ramos
[2] Ilias

دهدشتی شاهرخ (۱۳۹۸) به تحلیل متن کاوی نشریه برنامه ریزی و توسعه گردشگری با استفاده از تکنیک متن کاوی پرداخته است. در این پژوهش ساختار محتوایی مقالات بررسی و حوزه های مطالعاتی این فصلنامه مورد بررسی قرارگرفته اند .

تجزیه و تحلیل موضوعی مقالات منتشرشد کتابداری و اطلاع رسانی پزشکی در ایران با استفاده از فنون متن کاوی نیز توسط داستانی و همکاران (۱۳۹۹) مورد مطالعه قرار گرفت. این مطالعه با استفاده از رویکرد اکتشافی و توصیفی به تجزیه و تحلیل مقالات کتابداری و اطلاع رسانی منتشرشده در مجلات تخصصی این حوزه در ایران از سال ۱۳۷۶ تا ۱۳۹۸ با استفاده از فنون متن کاوی پرداخته است. جهت تعیین موضوعات منتشرشده، الگوریتم TF-IDF انتخاب گردیده است. جهت شناسایی مهمترین واژگان به کار رفته در مقالات از الگوریتم وزن دهی و از زبان برنامه نویسی پایتون نیز جهت اجرای الگوریتمهای متن کاوی استفاده شده است.

نتایج پژوهش حوزه های موضوعی مقالات را شناسایی نمود. توسعه و روند موضوعی حوزه علم اطلاعات و دانش نشاسی نیز بر اساس مدل موضوعی LDA توسط باغ محمد و همکاران (۱۳۹۹۹) مورد بررسی قرار گرفته است. به منظور دستیابی به اهداف پژوهش، داده ها از سال ۲۰۰۸ تا ۲۰۱۹ از پایگاه اسکوپوس استخراج شدند. سپس با استفاده از الگوریتمهای متن کاوی و به طور خاص الگوریتم مدل سازی موضوعی LDA با استفاده از نرم افزار R مورد تحلیل قرار گرفتند. نتایج نشان داد که پژوهشهای موضوعی رشتهٔ علم اطلاعات و دانش شناسی در ایران همگام با رشد فناور ی ها و موضو عهای جهانی توسعه یافته است. قناد و همکاران (۱۴۰۲) به کاربست فنون متن کاوی در تحلیل جریان موضوعی مقالات منتشر شده در مجلات حسابداری ایران پرداختند. در این پژوهش به‌منظور شناسایی مهمترین واژگان بکاررفته در مقالات از الگوریتم وزن‌دهیTF-IDF و جهت کاربست الگوریتم‌های متن‌کاوی از زبان برنامه‌نویسی پایتون استفاده شده است. نتایج نشان داد که توجه به موضوعاتی نظیر فناوری و سیستم‌های اطلاعاتی حسابداری، حسابداری بخش عمومی و آموزش و پژوهش حسابداری در مقایسه با سایر حوزه‌های پژوهشی کمتر بوده است

در مجموع نیزیافته های مطالعه آرامو (۲۰۱۸) نشان داده است تخصص در مقابل تنوع در پژوهش‌هاپژوهش‌ها علمی نتایج بهتری را به دنبال دارد. همچنین نتایج مطالعه حری (۱۳۷۴) حاکی از انست که بین مجلات و مقالات منتشره حوزه‌های تخصصی و ارتقا علمی متخصصان رابطه مثبت وجود دارد. میزان تخصص گرایی و تنوع موضوعی در مقالات نشریات با استفاده از روشهای متن کاوی در برخی نشریات مورد مطالعه قرار گرفته است.

با توجه به مرور پیشینه ها، برای بررسی متون علمی، شاهد حجم عظیمی از داده هستیم که برای بررسی ساختار و تحلیل روند موضوعی آنها استفاده از روش های سنتی و دستی، امکان پذیر نبوده یا با خطا مواجه خواهد بود. بنابراین همگام با توسعه فناوری های پردازش متن و یادگیری ماشین در تحلیل داده های بزرگ رویکردها ی متن کاوی برا ی شناسایی روند موضوعی پژوهشها مورد استفاده قرار گرفته است (لی و همکاران،۲۰۱۸)؛ اما همانگونه که مشاهده شد غالب پژوهشها به بررسی و مرور محتوای مقالات نشریات به صورت خاص پرداخته و نگرش جامعی در این خصوص وجود ندارد. بنابراین پژوهش حاضر بر آنست که میزان همخوانی دامنه موضوعی نشریات معتبر وزارت علوم، پژوهش‌هاپژوهش‌ها و فناوری در حوزه های شش‌گانه را با محتوای مقالات آنها و بر اساس فنون متن کاوی مورد تجزیه و تحلیل قرار دهد.

فصل سوم

روش‌شناسی پژوهش

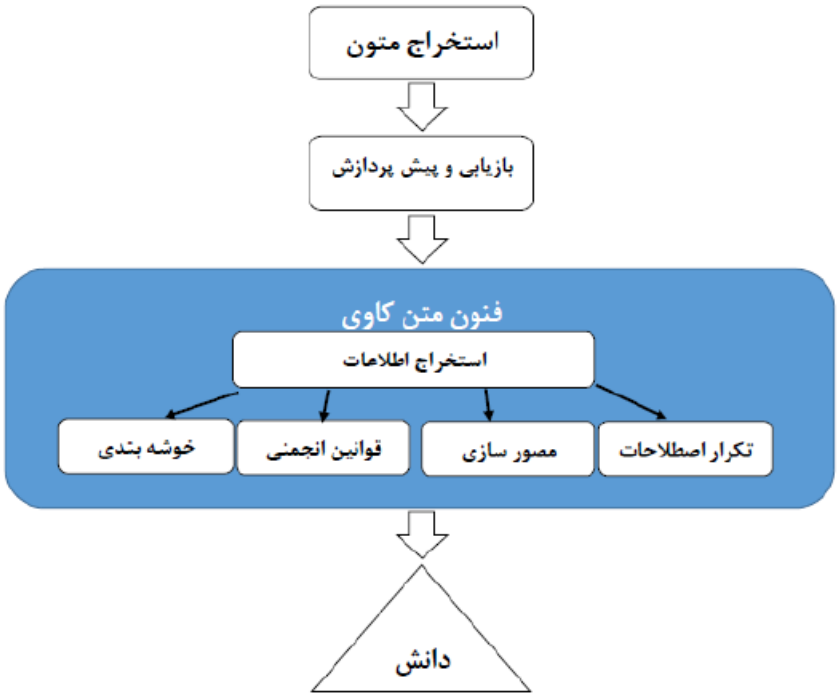
۳-۱. مقدمه

در این فصل روش‌شناسی پژوهش شامل نوع، روش و رویکرد، جامعه پژوهش، روش‌های گردآوری و تجزیه و تحلیل داده‌ها، مراحل اجرای پژوهش تبیین‌شده است.

۳-۲. روش پژوهش

پژوهش از نوع کاربردی و به روش متن کاوی با رویکرد تحلیلی انجام شده است. کاربردی بودن این پژوهش، ازآن‌جهت است که وضع موجود قلمرو پژوهش را بررسی کرده و از نتایج به‌دست‌آمده می توان در تصمیم‌گیری‌ها و سیاست‌گذاری‌ها و همچنین برنامه‌ریزی‌ها استفاده کرد. همچنین در این پژوهش با استفاده از رویکرد اکتشافی- توصیفی به تجزیه و تحلیل مقالات منتشر شده در نشریات معتبر وزارت علوم پرداخته شد. این مطالعه به دلیل استفاده از فنون متن کاوی یک مطالعه اکتشافی است. متن کاوی یک روش اکتشافی داده محور است که جهت یافتن الگوها و روندها در مجموعه های داده های بزرگ استفاده میشود.

روش متن کاوی به کار رفته در این پژوهش برگرفته از چارچوب طراحی شده توسط ژانگ و چن و توسعه یافته توسط سالوم و همکاران است که شامل پیش پردازش متن، عملیات متن کاوری و پسا پردازش است. پیش پردازش متن شامل عملیات زیر است: انتخاب داده ها، دسته بندی، استخراج ویژگی، نرمال سازی، حذف کلمات زائد. دومین مرحله با فنون مختلف متن کاوی مانند خوشه بندی، بصری سازی و تکرار واژگان سروکار دارد. در خلال مرحله سوم تغییر و تبدیلهایی روی داده ها (مقالات علمی) از طریق توابع متن کاوی مانند ارزیابی و انتخاب دانش؛ تجزیه و تحلیل و بصری سازی دانش صورت میگیرد.



۳-۳. جامعه پژوهش

در ابتدا فهرستی از نشریات فارسی مصوب وزارت علوم، تحققات و فناوری مشتمل بر ۱۱۵۸عنوان نشریه دارای رتبه طبق پرتال نشریات[۱۵۶] (ابان ۱۴۰۰) تهیه و سپس نمونه گیری به روش طبقه ای- تناسبی از بین شش حوزه موضوعی سطح کلان از پیش تعیین شده توسط وزارت عتف شامل علوم انسانی، فنی و مهندسی، علوم پایه، هنر و معماری، کشاورزی و منابع طبیعی و دامپزشکی انجام شد.

در جدول ۱ تعداد تمامی مجلات حوزه های مورد مطالعه و نمونه پژوهش به تفکیک حوزه موضوعی نشریه درج شده است. باتوجه به محدودیت‌های زمانی، منابع انسانی و مالی امکان بررسی اهداف و دامنه موضوعی مقالات تمام نشریات در پژوهش حاضر وجود ندارد. بنابراین در این پژوهش بایستی حجم نمونه مشخص گردد. در همین راستا، با مشورت گرفتن از مشاور آماری طرح جهت تعیین حجم نمونه، از نرم افزار PASS[۱۵۷] نسخه ۱۱ استفاده شد (هینتز[۱۵۸]، ۲۰۱۳). با در نظر گرفتن فاصله

156 journals.msrt.ir
157 - Power Analysis Sample Size
158 - Hintze

اطمینان ۰/۹۵، توان آزمون ۰/۸۰ و همبستگی مورد انتظار ۰/۲۰، حداقل حجم نمونه مورد نیاز توسط نرم افزار، ۱۱۶ عنوان نشریه تعیین شد. حجم نمونه مورد نظر با استفاده از روش نمونه‌گیری تصادفی طبقه‌ای انتخاب شد به گونه‌ای که تناسب بین نمونه و جامعه در هر یک از شش حوزه موضوعی سطح کلان تعیین شده توسط وزارت عنف رعایت شود.

توزیع فراوانی نشریات و تعداد نمونه پژوهش در حوزه های موضوعی شش گانه وزارت عتف

ردیف	حوزه موضوعی	تعداد نشریه فارسی در حوزه (عنوان)	تعداد نمونه (عنوان)
۱	علوم انسانی	۷۲۵	۷۲
۲	علوم کشاورزی و منابع طبیعی	۱۵۸	۱۶
۳	فنی و مهندسی	۱۳۴	۱۳
۴	علوم پایه	۸۸	۹
۵	هنر و معماری	۴۲	۴
۶	دامپزشکی	۱۱	۲
جمع		۱۱۵۸	۱۱۶

۴-۳. گردآوری داده ها

نکته ای که در این نوع مطالعات بایستی مورد توجه قرار گیرد و در حجم نمونه لحاظ شود؛ آن است که نمونه پژوهش در واقع شامل اطلاعات بخش حوزه موضوعی مندرج در وب سایت رسمی نشریه از یک سو و عنوان، کلیدواژه و چکیده مقالات دوره سه ساله هر یک از نشریات از سوی دیگر است. بنابراین در مرحله اول، وب سایت ۱۱۶ نشریه به منظور تعیین حوزه موضوعی درج و اعلام شده توسط نشریه بررسی و اطلاعات لازم گردآوری شد. در ادامه نشریات نمونه، در موتور

جستجوی گوگل به منظور بازیابی وب سایت اصلی مجله مورد جستجو قرار گرفت. سپس عنوان، چکیده و واژه های کلیدی مقالات هر نشریه طی یک بازه زمانی سه ساله استخراج شد که بالغ بر ۱۵۵۰۰۰ عنوان مقاله را دربرگرفت و در قالب بیب اکسل آماده سازی شد.

۵-۳. تجزیه و تحلیل داده‌ها

تجزیه‌وتحلیل داده‌ها و انجام عملیات متن کاوی در این پژوهش، شامل سه مرحله اصلی پیش پردازش،[۱۵۹] انجام عملیات و فنون متن کاوی و در پایان تحلیل و نتایج اصل از متن کاوی است. عملیات پاک‌سازی داده‌ها[۱۶۰] روی داده‌های متنی اجرا شد. پاک‌سازی داده‌ها، کیفیت داده‌ها، اعتبار الگوها و روابط استخراج‌شده را افزایش می دهد. پاک‌سازی داده‌ها، تنها داده‌های متنی مورد نیاز مرتبط را نگه می‌دارد (ایندوخیا و دامرو[۱۶۱]، ۲۰۱۰).

الگوریتمهای مدلسازی موضوعی با استفاده از تکنیکهای مختلف تلاش میکنند تا موضوعات خوشه بندی شده را تحت عنوان یک موضوع ارائه دهند. مدل توزیعی مانند فضای برداری، آنالیز پنهان مفهومی ، آنالیز پنهان مفهومی احتمالی و تخصیص دیریکله پنهان میتواند معنای کلمات را از متن با استفاده از روشهای آماری استخراج کند (زمانی و همکاران، ۱۳۹۳). به منظور انجام مدلسازی موضوعی در متن، نیاز به پردازش زبان میباشد. ابزارهای استاندارد پیش‌پردازش و نرمالسازی ایجاد شده برای متون زبان فارسی به صورت رایگان منتشر نشده یا دقت مناسب را ندارند (کامیابی و همکاران، ۱۳۹۷. به دلیل نزدیکی دبیره زبان فارسی با عربی، همواره در نگارش تعدادی از حرف‌ها مشکل نویسه‌های عربی معادل وجود دارد. از جملهٔ آنها می‌توان به حروف «ک»، «ی»، همزه و... اشاره کرد. علاوه بر این، حذف نویسهٔ «ـ»، تشدید، تنوین و موارد مشابه (کامیابی و همکاران، ۱۳۹۷؛ سراجی[۱۶۲]، ۲۰۱۰) ایست واژه‌ها[۱۶۳] از جمله حروف اضافه، بسیاری از قیدها، حروف ربط و افعال از جمله

[159] Preprocessing
[160] data cleaning
[161] Indurkhya and Damerau
[162] Seraji
[163] Stop word

اقدام‌های لازم قبل از شروع پردازش متن است. سپس تنوع ریخت‌شناسی واژه‌ها در زبان فارسی طبق آخرین یافته‌ها از متون استخراج و مورد استفاده قرار گرفت (هماوند و همکاران ۱۳۹۷؛ پرئی و حمیدی ۱۳۹۶؛ ستوده و هنرجویان ۱۳۹۱). که این پردازش در هر سطح، نیازمند دانش، منابع و پیکره های مورد نیاز آن سطح و سطوح پایینتر است. در دسترس بودن منابع و دانش برای انجام تحقیق در حیطه ی پردازش زبان طبیعی از جمله چالشهای پردازش زبان طبیعی است. در زبان فارسی، این مشکلات و چالشها به مراتب بیشتر هم میشود؛ چرا که زبان فارسی به صورت ماهوی از پیچیدگیهای بیشتری برخوردار است و پژوهشهای به نسبت بسیار کمتری روی آن انجام گرفته است .پیوسته بودن نویسه ها، شباهت کاراکترها، کاراکترهای همآوا و وجود برخی اسامی مرکب دو یا چند کلمه ای موجب افزایش خطا در متون فارسی میشود. ناهماهنگیهای گوناگونی که در نگارش خط فارسی دیده میشود، همچنین در نگارش رایانه ای متن فارسی، قالبها، ابزارها، سیستم عاملهای گوناگون و روشهای گوناگون کد کردن نوشته ی فارسی دیده میشود .ویژگیهای منحصربه فرد خط فارسی موجب بروز چالشهایی برای فرآیند خطایابی و تصحیح خطا میشود که برای دیگر زبانهایی که این خصوصیات و استثناءها را ندارند، مطرح نیست . عدم دسترسی به پیکره ها، دانش مورد نیاز و برخی چالشهای وابسته به رسم الخط زبان فارسی موجب میشود، پژوهش در این زمینه با مشکلات عدیده ای روبرو شده و پیشرفت در این زمینه به کندی پیش رود

در این مرحله، پیش پردازش داده ها با استفاده از برنامه نویسی پایتون بمنظور جداسازی واژه ها، یکدستی و نرمالسازی داده ها و به شرح زیر صورت گرفت.

- حذف نویسه‌های غیرمهم مانند فضاهای خالی اضافی، تگ‌های قالب‌بندی متن، حذف نویسه‌های غیرالفبایی شامل حذف علائم نگارشی و اعداد از متن است (بانکس[۱۶۴] و همکاران، ۲۰۱۸). بنابراین، این روش برای تمرکز روی کلمات و عبارات موجود در متن انجام شده است.

[164] Banks

- شکستن اجزای متن[۱۶۵] به کلمات و واژه‌ها، شکستن اجزای متن به فرایند جداسازی و شکستن هر کلمه در سند براساس واحدهایی با معنی، شکستن یا تقسیم‌بندی متن گفته می‌شود (ویجایارانی[۱۶۶] و همکاران، ۲۰۱۵) و یکی از فرآیندهای ضروری پیش‌پردازش متن در متن‌کاوی است (اشمیدل[۱۶۷] و همکاران، ۲۰۱۹). که در این پژوهش داده‌های متنی براساس کلمات[۱۶۸] جداسازی شده است.
- نرمالسازی و یکدست سازی حروف مانند حروف«ی»، «ک»، همزه و... جهت یکدست‌سازی متن
- حذف ایست‌واژه‌ها: به‌منظور بازیابی یا تجزیه‌وتحلیل درست، حروف یا واژه‌هایی که ارزشی معنایی ندارند، با استفاده از سیاهه ایست‌واژه‌ها حذف می‌شوند (بانکس و همکاران، ۲۰۱۸). ایست‌واژه‌ها واژه‌هایی هستند که محتوای اطلاعاتی کم‌ارزشی یا بی‌ارزشی دارند و به معنادار کردن متن کمکی نخواهند کرد.. افزون بر ایست‌واژه‌ها، داده‌های متنی بررسی شد و واژگان دیگری که در متن تکرارشده و معنی خاصی در متن نداشتند به سیاهه ایست‌واژه‌ها اضافه و از متن حذف شدند.
- در این پژوهش از الگوریتم وزن‌دهی واژگان TF-IDF استفاده گردید.

در مرحله بعدی تعیین فراوانی و وزن‌دهی واژگان، اجرای الگوریتم شباهت سنجی با استفاده از مدل فضای برداری انجام شد. بدین ترتیب که عبارات کلیدی مندرج در بخش دامنه موضوعی نشریه و همچنین عبارات کلیدی مستخرج از محتوای مقالات به یک سری بردار عددی تبدیل شدند. این بردارها دارای n مولفه بوده و وزندهی به آنها بر اساس تعداد تکرار کلمات لحاظ شده است. در نهایت با استخراج دانش از متون و تفسیر آن به پایان رسید.

- مقدار TF-IDF به‌تناسب تعداد تکرار واژه در سند افزایش می‌یابد و توسط تعداد اسنادی که در مجموعه هستند، متعادل می‌گردد (اوکوهارا[۱۶۹] و همکاران، ۲۰۱۸؛ وی‌جیرانی و جنانی، ۲۰۱۶؛ وو[۱۷۰] و همکاران، ۲۰۰۸).

---

165 Tokenization
166 Vijayarani
167 Schmiedel
168 Word tokenize
169 Okuhara
170 Wu

فصل چهارم

تجزیه‌وتحلیل داده‌ها

۴-۱. مقدمه

در این فصل یافته های حاصل از تحلیل داده های پژوهش توصیف شده و با ترتیبی منطقی ارائه گردیده است. به این صورت که داده های تحلیل شده در قالب جدولها و نمودارها گزارش و توصیف های مربوط به هرکدام ارائه گردیده است.

۴-۲. استخراج و آماده‌سازی دامنه موضوعی نشریات

به منظور رسیدن به هدف پژوهش ابتدا دامنه موضوعی نشریات استخراج و عبارات کلیدی پس از پیش پردازش ماشینی توسط پژوهشگر نیز کنترل و استخراج شد. به دلیل حجم زیاد داده ها جدول شمایی از فرآیند را نشان می دهد.

استخراج عبارت کلیدی از دامنه و اهداف موضوعی درج شده در وبگاه هر نشریه

ردیف	عنوان نشریه	حوزه موضوعی مندرج در وبگاه	عبارات کلیدی
۱	پردازش علائم و داده ها	- انتشار دانش علمی و تخصصی در زمینه‌های پردازش صوت، تصویر، متن، رمز، امنیت اطّلاعات و مهندسی پزشکی - توسعهٔ پل ارتباطی بین پژوهشگران و پژوهشگران کشور - کمک به ارتقای سطح دانش در زمینه‌های مورد بحث دوفصل نامهٔ پردازش علایم و داده‌ها - تلاش در جهت رفع نیازهای علمی، پژوهشی و پژوهش‌هاپژوهش‌های دانشجویان، پژوهش‌گران و	-پردازش صوت - پردازش تصویر - پردازش متن - پردازش رمز - امنیت اطّلاعات - مهندسی پزشکی - پردازش علایم و داده-ها

		پژوهشگران مراکز پژوهشی، پژوهش‌هاپژوهش‌های و دانشگاه	
۲	مطالعات الگوی پیشرفت اسلامی ایرانی	- شناخت و معرفی و ارائه تحلیل، تبیین نظری و الگوههای عملی و همچنین نگرشهای علمی در رابطه با موضوع الگوی پیشرفت - انتشار مقالات تخصصی علمی – پژوهشی در گرایش های مرتبط و ارائه نتایج پژوهش‌هاپژوهش‌ها نظری و عملی پژوهشگران. - فراهم کردن زمینه های گسترش دانش مطالعات الگوی پیشرفت اسلامی ایرانی در وجوه مختلف فرهنگی،اقتصادی و... منطبق با نگرشهای تبین کننده گفتمان انقلاب اسلامی، از طریق انتشار آخرین یافته‌های پژوهشی، تحلیل جدیدترین نظریه‌ها و معرفی آخرین نظر و پژوهش‌های منتشرشده مرتبط با موضوع	-الگوی پیشرفت اسلام - پژوهش‌هاپژوهش‌ها نظری - تخقیقات عملی - الگوی پیشرفت فرهنگی - الگوی پیشرفت اقتصادی -گفتمان انقلاب اسلامی
۳	پژوهشهای پولی- بانکی	انتشار مطالعات و پژوهش‌هاپژوهش‌ها علمی اقتصادی در زمینه‌های اقتصاد کلان مرتبط با سیاست پولی، ارزی، بانکی و اعتباری و اقتصاد خرد مرتبط با بانکداری است .الزامی بر ارتباط مستقیم مطالعات با اقتصاد ایران وجود ندارد، اما انتظار می‌رود یافته‌های پژوهش‌ها به	-اقتصاد کلان - سیاست پولی - سیاست ارزی - سیاست بانکی و اعتباری

		درک بهتر مسائل مبتلابه اقتصاد ایران به‌خصوص در حوزه پول و بانک کمک کند. بر این اساس، خط مشی مجله ترغیب پژوهشگران علاقه‌مند به مباحث اقتصادی مرتبط با حوزه‌های پولی و بانکی برای انجام مطالعات نظری و کاربردی در این زمینه‌ها و نیز آشنا کردن جامعه علمی کشور با پژوهش‌هاپژوهش‌ها ناب و عمیق اقتصادی در زمینه‌های مرتبط با نظام بانکی اعم از بانکداری مرکزی و تجاری است. نشر یافته‌ها و نظریه‌های جدید در زمینه‌های اقتصاد پولی و بانکی و فراهم آوردن زمینه‌های تبادل نظر در این حوزه از دیگر اهداف مجله تعریف شده است.	- اقتصاد خرد - بانکداری - اقتصاد ایران - پول و بانک - نظام بانکی - بانکداری مرکزی - بانکداری تجاری - اقتصاد پولی و بانکی

۴-۳. استخراج و آماده‌سازی عناصر مورد نیاز از محتوای مقالات نشریات

در این مرحله اطلاعات بالغ بر ۱۵۵۰۰۰ عنوان مقاله از نشریات مورد بررسی استخراج گردید. جدول ۲ شمایی از این فرآیند را نشان می دهد.

جدول ۲- استخراج عناصر مقالات از هر نشریه

id	key_word	text	label
1	طراحي مدل يكپارچه توسعه سطح نوآوري و تجاري سازي ◀ نوآوري ◀ تجاري سازي ◀ دانش بنيان ◀ نظريه داده		ابتكار و خلاقيت در علوم انساني
2	تاثير خلاقيت و شيوه هاي نوآورانه دانش پايه سي در پيش ◀ دانش پايه سي ◀ خلاقيت و نوآورانه ◀ پيشگيري و كنت		ابتكار و خلاقيت در علوم انساني
3	رابطه هوش‌هاي چندگانه با ويژگي شخصيتي كارآفريني و ◀ هوش چند گانه ◀ ويژگي شخصيتي كار آفريني ◀ خلا		ابتكار و خلاقيت در علوم انساني
4	طراحي مدل پرورش تفكر استراتژيك و خلاق در مديران ◀ پرورش تفكر استراتژيك ◀ روانشناسي استراتژيك		ابتكار و خلاقيت در علوم انساني
5	اثربخشي آموزش تفكر انتقادي بر نگرش به خلاقيت و نشا ◀ تفكر انتقادي ◀ نگرش به خلاقيت ◀ نشاط ذهني		ابتكار و خلاقيت در علوم انساني
6	نوآوري و خلاقيت در اجراي اقدامات مديريت منابع انساني ◀ مديريت منابع انساني در شركت‌هاي دانش‌بنيان ◀ مدير		ابتكار و خلاقيت در علوم انساني
7	الگوي كاربردي پرورش خلاقيت كودكان از طريق موسيقي ◀ الگوي پرورش خلاقيت ◀ پرورش خلاقيت از طريق آم		ابتكار و خلاقيت در علوم انساني
8	سنجش پتانسيل تكنولوژي آموزشي نو آورانه ي نگارش مش ◀ نگارش مشاركت الكترونيكي ◀ مذاكره و انتقال معنا و		ابتكار و خلاقيت در علوم انساني
9	تأثيرفناوري هاي هوشمندسازي مدارس برخلاقيت دانش آ ◀ خلاقيت ◀ تورنس ◀ فناوري آموزشي ◀ هوشمندسا		ابتكار و خلاقيت در علوم انساني
10	تاثير يادگيري مشاركتي برپرورش خلاقيت دانش آموزان ◀ يادگيري مشاركتي ◀ خلاقيت ◀ جغرافيا		ابتكار و خلاقيت در علوم انساني
11	طراحي مدل نوين خط مشي گذاري نظارت بر تجارت فرامرزي ◀ خط‌مشي‌گذاري ◀ نظارت بر تجارت فرامرزي ◀ پنجر		ابتكار و خلاقيت در علوم انساني
12	تجزيه تحليل نقش واسطه‌اي بازخورد كار از سرپرست د ◀ استرس ◀ بازخورد ◀ خلاقيت ◀ كاركنان		ابتكار و خلاقيت در علوم انساني
13	نوآوري در جذب منابع مالي: تأملي بر چالش هاي مالي مفو ◀ نوآوري در جذب منابع مالي ◀ چالش مالي مدرسه ◀ موا		ابتكار و خلاقيت در علوم انساني
14	بررسي مؤلفه هاي مؤثر بر مصرف رسانه اي جوانان ته ◀ اعتماد رسانه اي ◀ استفاده و رضايتمندي ◀ مخاطبان		ابتكار و خلاقيت در علوم انساني
15	بررسي محتواي كتاب رياضي پنج پايه دهم دوره متوسطه بتحليل محتوا ◀ محتوا ◀ خلاقيت ◀ عوامل خلاقيت گي		ابتكار و خلاقيت در علوم انساني
16	بررسي ويژگي هاي روانسنجي مقياس دامنه هاي خلاقيت ◀ روانسنجي ◀ مقياس ◀ خلاقيت ◀ كاهن		ابتكار و خلاقيت در علوم انساني
17	بررسي اثربخشي تئاتر پداگوژيك بر ارتقاء سطح هوش هي ◀ تئاتر پداگوژيك ◀ هوش هيجاني ◀ كودكان		ابتكار و خلاقيت در علوم انساني
18	تبيين علّي خلاقيت: نقش سبك هاي دلبستگي و هيجانات مث ◀ خلاقيت ◀ سبك دلبستگي ◀ هيجان مثبت ◀ هيجان م		ابتكار و خلاقيت در علوم انساني
19	ساختمان دانشكده معماري، زمينه پرورش خلاقيت دانشجو ◀ خلاقيت ◀ معماري دانشگاه ◀ آموزش ◀ روانشناس		ابتكار و خلاقيت در علوم انساني
20	اجراي استراتژي و عملكرد با توجه به نقش تحليل‌گر هاي ◀ اجراي استراتژي ◀ خلاقيت ◀ عملكرد ◀ مديريت ا		ابتكار و خلاقيت در علوم انساني

داده های بدست امده از این مرحله مورد پیش پردازش و اماده سازی برای استفاده توسط ماشین قرار گرفت (جدول ۳).

جدول ۳- یکدست سازی و تعیین فراوانی عبارات

index	key_word
خلاقيت	51
نوآوري	8
دانشگاه آزاد اسلامي	5
دانش‌آموزان	3
نوآوري سازماني	3
تفكر خلاق	3
تفكر انتقادي	3
رفتار نوآورانه	3
دانش آموزان	3
كاركنان	3
معلمان	2

۴-۴. خوشه بندی داده‌ها

در این مرحله خوشه‌بندی داده‌ها انجام شد (جدول ۴).

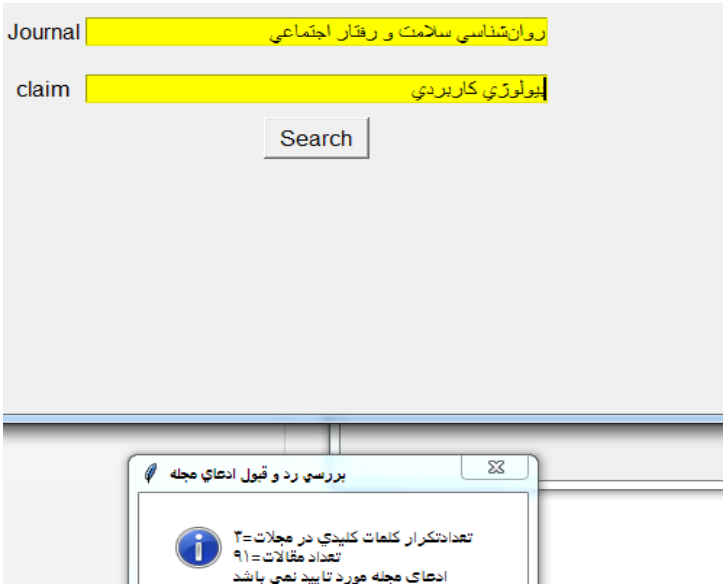
جدول ۴- خوشه بندی داده ها

اجتماعي	آموزان	دانشگاه	توآوري
هوش اجتماعي	دانش آموزان	دانشگاه آزاد اسلامي	توآوري
گزارشگري اجتماعي	دانش‌آموزان تيزهوش تحصيلي	دانشگاه هاي دولتي	توآوري سازماني
افشا مسئوليت اجتماعي	دانش آموزان ابتدائي	دانشگاه فرهنگيان	توآوري دار
كاهش ترس مبتني بر شبكه هاي اجتماعي	دانش آموزان دختر	دانشگاه تربيتان	توآوري منطقه اي
	دانش آموزان تيزهوش	توسعه ي كارآفريني دانشگاهي	توآوري در معماري
پيوند شبكه اجتماعي	دانش‌آموزان دوره متوسطه اول منطقه ۳ شهر تهران	دارايگردادن دانشگاهي	توآوري نظام يافته
شبكه هاي اجتماعي		معماري دانشگاه	توآوري راهبردي
مهارت هاي اجتماعي		حمايت خانواده و دانشگاه	توآوري آموزش عالي
بعد اجتماعي – فرهنگي			توآوري وب سايت ها
			توآوري كاري
			خلاقيت و توآوري
			توآوري اجتماعي
			تعالگر توآوري
			توآوري در اقدامات مديريت منابع انساني
			توآوري در جذب منابع مالي

۵-۴. شباهت‌سنجی

در مرحله پایانی شباهت‌سنجی انجام شد (عکس ۱).

عکس ۱- شباهت سنجی حوزه موضوعی نشریه با محتوای مقالات



۵-۵ یافته ها

میزان همخوانی یا شباهت موضوعات مندرج در وبگاه هر نشریه و محتوای مقالات به شرح جداول ۷ تا ۱۲ محاسبه گردید.

جدول ۷- میزان همخوانی محتوای مقالات نشریات حوزه دامپزشکی با حوزه موضوعی درج شده در وبگاه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	دامپزشکی	دامپزشکی	التیام	۶۰
۲	دامپزشکی	دامپزشکی	بهداشت مواد غذایی	۷۰

همانگونه که مشاهده می‌شود در نشریات مورد بررسی حوزه موضوعی دامپزشکی، همخوانی به میزان ۶۰ و ۷۰ درصد رعایت شده است.

جدول ۸- میزان همخوانی محتوای مقالات نشریات حوزه علوم انسانی با حوزه موضوعی درج شده در وبگاه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	علوم انسانی	علوم تربیتی	ابتکار و خلاقیت در علوم انسانی	۹۵
۲	علوم انسانی	جغرافیا	اقتصاد فضا و توسعه روستایی	۹۵
۳	علوم انسانی	روانشناسی	پژوهش های روانشناسی اجتماعی	۹۵
۴	علوم انسانی	علوم اجتماعی	اعتیاد پژوهی	۹۵
۵	علوم انسانی	علوم سیاسی	پژوهش های انقلاب اسلامی	۹۵
۶	علوم انسانی	مدیریت	اندیشه آماد	۹۵
۷	علوم انسانی	علوم قرآن و حدیث	آموزه های تربیتی در قرآن و حدیث	۸۰
۸	علوم انسانی	زبان و ادبیات	ادبیات عرفانی	۷۰
۹	علوم انسانی	اخلاق	پژوهش های اخلاقی	۷۰
۱۰	علوم انسانی	ادیان، مذاهب و عرفان	اندیشه نوین دینی	۷۰
۱۱	علوم انسانی	تاریخ	پژوهش های باستان‌شناسی ایران	۷۰
۱۲	علوم انسانی	تربیت بدنی	پژوهش در توانبخشی ورزشی	۷۰
۱۳	علوم انسانی	جغرافیا	پژوهش های جغرافیای طبیعی	۷۰

۱۴	علوم انسانی	جغرافیا	پژوهش های ژئومورفولوژی کمی	۷۰
۱۵	علوم انسانی	حسابداری	بررسی های حسابداری و حسابرسی	۷۰
۱۶	علوم انسانی	روانشناسی	اندازه گیری تربیتی	۷۰
۱۷	علوم انسانی	زبان و ادبیات	پژوهش های ادبیات تطبیقی	۷۰
۱۸	علوم انسانی	علوم اجتماعی	پژوهش های اطلاعاتی و جنایی	۷۰
۱۹	علوم انسانی	علوم تربیتی	آموزش محیط زیست و توسعه پایدار	۷۰
۲۰	علوم انسانی	علوم سیاسی	پژوهش های روابط بین الملل	۷۰
۲۱	علوم انسانی	فقه و حقوق	آموزه های حقوق کیفری	۷۰
۲۲	علوم انسانی	فلسفه و کلام	پژوهش های علم و دین	۷۰
۲۳	علوم انسانی	مدیریت	آینده پژوهی ایران	۷۰
۲۴	علوم انسانی	اخلاق	اخلاق وحیانی	۶۰
۲۵	علوم انسانی	ادیان، مذاهب و عرفان	ادیان و عرفان	۶۰
۲۶	علوم انسانی	ادیان، مذاهب و عرفان	پژوهش های ادیانی	۶۰
۲۷	علوم انسانی	اقتصاد	اقتصاد انرژی ایران	۶۰
۲۸	علوم انسانی	اقتصاد	اقتصاد مالی	۶۰
۲۹	علوم انسانی	اقتصاد	اقتصاد و برنامه ریزی شهری	۶۰
۳۰	علوم انسانی	اقتصاد	اقتصاد و تجارت نوین	۶۰
۳۱	علوم انسانی	اقتصاد	برنامه ریزی و بودجه	۶۰
۳۲	علوم انسانی	تاریخ	پژوهش های تاریخی ایران و اسلام	۶۰
۳۳	علوم انسانی	تربیت بدنی	پژوهش در طب ورزشی و فناوری	۶۰
۳۴	علوم انسانی	تربیت بدنی	پژوهش در مدیریت ورزشی و رفتار حرکتی	۶۰
۳۵	علوم انسانی	جغرافیا	آمایش جغرافیایی فضا	۶۰
۳۶	علوم انسانی	جغرافیا	برنامه ریزی توسعه کالبدی	۶۰
۳۷	علوم انسانی	جغرافیا	پژوهش های اقلیم شناسی	۶۰
۳۸	علوم انسانی	حسابداری	پژوهش های حسابداری مالی و حسابرسی	۶۰
۳۹	علوم انسانی	روانشناسی	پژوهش در سلامت روانشناختی	۶۰

۴۰	علوم انسانی	زبان و ادبیات	ادب فارسی	۶۰
۴۱	علوم انسانی	زبان و ادبیات	ادبیات پارسی معاصر	۶۰
۴۲	علوم انسانی	زبان و ادبیات	ادبیات تطبیقی	۶۰
۴۳	علوم انسانی	زبان و ادبیات	ادبیات و زبانهای محلی ایران زمین	۶۰
۴۴	علوم انسانی	زبان و ادبیات	پژوهش های زبان شناسی	۶۰
۴۵	علوم انسانی	علم اطلاعات و دانش شناسی	بازیابی دانش و نظام های معنایی	۶۰
۴۶	علوم انسانی	علوم اجتماعی	بررسی مسایل اجتماعی ایران	۶۰
۴۷	علوم انسانی	علوم اجتماعی	برنامه ریزی رفاه و توسعه اجتماعی	۶۰
۴۸	علوم انسانی	علوم اجتماعی	پژوهش های انسان شناسی ایران	۶۰
۴۹	علوم انسانی	علوم تربیتی	الگوی پیشرفت اسلامی ایرانی	۶۰
۵۰	علوم انسانی	علوم تربیتی	اندیشه های نوین تربیتی	۶۰
۵۱	علوم انسانی	علوم تربیتی	پژوهش در مسائل تعلیم و تربیت	۶۰
۵۲	علوم انسانی	علوم تربیتی	پژوهش های برنامه درسی	۶۰
۵۳	علوم انسانی	علوم سیاسی	پژوهش سیاست نظری	۶۰
۵۴	علوم انسانی	علوم سیاسی	پژوهش های سیاست اسلامی	۶۰
۵۵	علوم انسانی	علوم قرآن و حدیث	پژوهش دینی	۶۰
۵۶	علوم انسانی	علوم قرآن و حدیث	پژوهش های تفسیر تطبیقی	۶۰
۵۷	علوم انسانی	فقه و حقوق	پژوهش حقوق خصوصی	۶۰
۵۸	علوم انسانی	فقه و حقوق	پژوهش حقوق کیفری	۶۰
۵۹	علوم انسانی	فلسفه و کلام	اندیشه دینی	۶۰
۶۰	علوم انسانی	فلسفه و کلام	پژوهش های اعتقادی کلامی	۶۰
۶۱	علوم انسانی	مالی	بورس اوراق بهادار	۶۰
۶۲	علوم انسانی	مدیریت	اندیشه مدیریت راهبردی	۶۰
۶۳	علوم انسانی	مدیریت	آینده پژوهی مدیریت	۶۰
۶۴	علوم انسانی	اقتصاد	اقتصاد باثبات	۴۰
۶۵	علوم انسانی	تاریخ	پژوهش های تاریخی	۴۰
۶۶	علوم انسانی	حسابداری	پژوهش های حسابداری مالی	۴۰

۶۷	علوم انسانی	روانشناسی	پژوهش های روانشناسی بالینی و مشاوره	۴۰
۶۸	علوم انسانی	زبان و ادبیات	ادبیات عرفانی و اسطوره شناختی	۴۰
۶۹	علوم انسانی	علم اطلاعات و دانش شناسی	آینه میراث	۴۰
۷۰	علوم انسانی	علوم تربیتی	برنامه ریزی درسی نظریه و عمل	۴۰
۷۱	علوم انسانی	علوم تربیتی	پژوهش در یادگیری آموزشگاهی و مجازی	۴۰
۷۲	علوم انسانی	فلسفه و کلام	آینه معرفت	۴۰

همانگونه که مشاهده می‌شود در نشریات مورد بررسی حوزه علوم انسانی، همخوانی موضوعی در ۸ درصد نشریات ۹۵ درصد و پس از آن به ترتیب در ۱ درصد نشریات ۸۰ درصد، در ۲۲ درصد نشریات ۷۰ درصد، در ۵۶ درصد نشریات ۶۰ درصد و در ۱۳ درصد نشریات ۴۰ درصد رعایت شده است.

جدول ۹- میزان همخوانی محتوای مقالات نشریات حوزه علوم پایه با حوزه موضوعی درج شده در وبگاه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	علوم پایه	زمین شناسی	اکتشاف و تولید نفت و گاز	۸۰
۲	علوم پایه	زیست شناسی	اکوفیتوشیمی گیاهان دارویی	۷۰
۳	علوم پایه	محیط زیست	انسان و محیط زیست	۷۰
۴	علوم پایه	زمین شناسی	بلورشناسی و کانی شناسی ایران	۷۰
۵	علوم پایه	زیست شناسی	اقیانوس شناسی	۶۰
۶	علوم پایه	آمار	اندیشه آماری	۴۰
۷	علوم پایه	فیزیک	پژوهش سیستم های بس ذره ای	۴۰
۸	علوم پایه	ریاضی	پژوهش های ریاضی	۴۰
۹	علوم پایه	شیمی	پژوهش های شیمی(CR)	۴۰

همانگونه که مشاهده می‌شود در نشریات مورد بررسی حوزه علوم پایه، میزان همخوانی موضوعی در ۱۱ درصد نشریات ۸۰ درصد، در ۳۳ درصد نشریات ۷۰ درصد، در ۱۱ درصد نشریات ۶۰ درصد و در ۴۵ درصد نشریات ۴۰ درصد رعایت شده

است.

جدول ۱۰- میزان همخوانی محتوای مقالات نشریات حوزه فنی و مهندسی با حوزه موضوعی درج شده در وبگاه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	فنی و مهندسی	میان رشته ای	انرژی ایران	۷۰
۲	فنی و مهندسی	میان رشته ای	پدافند الکترونیکی و سایبری	۷۰
۳	فنی و مهندسی	مهندسی شیمی	پژوهش نفت	۷۰
۴	فنی و مهندسی	پلیمر	علوم و فناوری نساجی	۶۰
۵	فنی و مهندسی	مکانیک	انرژی های تجدیدپذیر و نو	۶۰
۶	فنی و مهندسی	پلیمر	بسپارش	۶۰
۷	فنی و مهندسی	برق و کامپیوتر	الکترو مغناطیس کاربردی	۴۰
۸	فنی و مهندسی	برق و کامپیوتر	پردازش سیگنال پیشرفته	۴۰
۹	فنی و مهندسی	برق و کامپیوتر	پردازش علائم و داده ها	۴۰
۱۰	فنی و مهندسی	عمران	پژوهش های زیرساخت های عمرانی	۴۰
۱۱	فنی و مهندسی	نفت	پژوهش های سیاستگذاری و برنامه ریزی انرژی	۴۰
۱۲	فنی و مهندسی	عمران	اساس	۴۰
۱۳	فنی و مهندسی	مواد و متالورژی	علوم و مهندسی سطح	۴۰

همانگونه که مشاهده می‌شود در نشریات مورد بررسی حوزه فنی و مهندسی، میزان همخوانی موضوعی در ۲۳ درصد نشریات ۷۰ درصد، در ۲۳ درصد نشریات ۶۰ درصد و در ۵۴ درصد نشریات ۴۰ درصد رعایت شده است.

جدول ۱۱- میزان همخوانی محتوای مقالات نشریات حوزه کشاورزی و منابع طبیعی با حوزه موضوعی مندرج در وبگاه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	کشاورزی و منابع طبیعی	زراعت	اکو فیزیولوژی گیاهی	۹۵

۲	کشاورزی و منابع طبیعی	زراعت	به زراعی کشاورزی	۹۵
۳	کشاورزی و منابع طبیعی	آب و خاک	پژوهش آب ایران	۹۵
۴	کشاورزی و منابع طبیعی	اقتصاد کشاورزی	اقتصاد کشاورزی و توسعه	۷۰
۵	کشاورزی و منابع طبیعی	اکولوژی	بوم شناسی کشاورزی	۷۰
۶	کشاورزی و منابع طبیعی	گیاهپزشکی	بیماریهای گیاهی	۷۰
۷	کشاورزی و منابع طبیعی	میان رشته‌ای	پژوهش های حبوبات ایران	۷۰
۸	کشاورزی و منابع طبیعی	محیط زیست	اکوبیولوژی تالاب	۶۰
۹	کشاورزی و منابع طبیعی	شیلات	آبزیان زینتی	۶۰
۱۰	کشاورزی و منابع طبیعی	گیاهپزشکی	آفات و بیماریهای گیاهی	۶۰
۱۱	کشاورزی و منابع طبیعی	شیلات	بهره برداری و پرورش آبزیان	۶۰
۱۲	کشاورزی و منابع طبیعی	جنگلداری	بوم شناسی جنگل های ایران	۶۰
۱۳	کشاورزی و منابع طبیعی	بیوتکنولوژی و ژنتیک گیاهی	پژوهش های ژنتیک گیاهی	۶۰
۱۴	کشاورزی و منابع طبیعی	آبخیزداری	پژوهش های آبخیزداری	۴۰
۱۵	کشاورزی و منابع طبیعی	صنایع غذایی	پژوهش های صنایع غذایی	۴۰

۱۶	کشاورزی و منابع طبیعی	اکولوژی	بوم شناسی کاربردی	۴۰

همانگونه که مشاهده می‌شود در نشریات مورد بررسی حوزه کشاورزی و منابع طبیعی، میزان همخوانی موضوعی در ۱۹ درصد نشریات ۹۵ درصد، در ۲۵ درصد نشریات ۷۰ درصد، در ۳۷ درصد نشریات ۶۰ درصد و در ۱۹ درصد نشریات ۴۰ درصد رعایت شده است.

جدول ۱۲- میزان همخوانی محتوای مقالات نشریات حوزه هنر و معماری با حوزه موضوعی مندرج در وبگاه نشریه

ردیف	حوزه موضوعی کلان	حوزه موضوعی میانی	عنوان نشریه	درصد همخوانی
۱	هنر و معماری	هنر و معماری	اثر	۶۰
۲	هنر و معماری	هنر و معماری	اندیشه معماری	۶۰
۳	هنر و معماری	هنر و معماری	باغ نظر	۶۰
۴	هنر و معماری	هنر و معماری	مبانی نظری هنرهای تجسمی	۶۰

همانگونه که مشاهده می‌شود میزان همخوانی موضوعی در همه نشریات مورد بررسی حوزه هنر و معماری به میزان ۶۰ درصد رعایت شده است.

جدول ۱۳- فراوانی نشریات بر اساس میزان همخوانی محتوای مقالات نشریه با حوزه موضوعی مندرج در وبگاه به تفکیک حوزه موضوعی

میزان همخوانی / حوزه موضوعی	٪۴۰	٪۶۰	٪۷۰	٪۸۰	٪۹۵
	فراوانی (تعداد نشریه)				
علوم انسانی	۹	۴۰	۱۶	۱	۶
عوم پایه	۴	۱	۳	۱	۰
فنی و مهندسی	۷	۳	۳	۰	۰
کشاورزی و منابع طبیعی	۳	۶	۴	۰	۳
دامپزشکی	۰	۱	۱	۰	۰
هنر و معماری	۰	۴	۰	۰	۰

همانگونه که مشاهده می شود از مجموع نشریات مورد بررسی ۸ درصد نشریات میزان همخوانی یا تطابق محتوای مقالات با دامنه موضوعی مندرج در وب سایت را به میزان ۹۵ درصد رعایت کرده بودند. کمترین میزان همخوانی ۴۰ درصد است که در ۲۰ درصد نشریات مورد بررسی وجود داشت. در این میان میزان همخوانی در ۴۷ درصد نشریات ۶۰ درصد، ۲۳ درصد نشریات ۷۰ درصد و در ۲ درصد نشریات به میزان ۸۰ درصد رعایت شده بود.

فصل پنجم

نتیجه‌گیری و ارائه پیشنهادها

۵-۱. مقدمه

در بخش نخست ابتدا نتایج حاصل از تجزیه‌وتحلیل داده‌ها ارائه می‌گردد، سپس پیشنهادهای پژوهش تبیین می‌شود.

۵-۲. بحث و نتیجه‌گیری

پژوهشگران، از مهمترین ارکان دانشی تأثیرگذار در نظام علم و فناوری محسوب می‌شوند. این افراد، علاوه بر فراهم آوردن موجبات توسعه و ترقی نظام آموزش عالی و ایجاد زمینه های لازم برای توسعه و پویایی علمی و فرهنگی جامعه، با ارائۀ دستاوردهای ارزشمند علمی و فناورانه خود، چرخ علم و فناوری را به گردش درمی‌آورند و سکان داران اصلی توسعه علم و فناوری محسوب می‌شوند.

باتوجه به اینکه نشریات علمی آخرین دستاوردهای علمی و افق های پژوهشی را در کوتاه ترین زمان ممکن منتشر می کنند و همچنین فرآیند داوری دقیقی را طی می کنند؛ یکی از مهمترین قالب های انتشار تولیدات علمی هستند. بنابراین انتظار می رود مقالات علمی منتشرشده در مجلات معتبر علمی، نماینده مناسبی از تولیدات پژوهشی یک رشته علمی باشند و تحلیل و مطالعه محتوای این مقالات بتواند در برنامه ریزی ها و سیاست گذار ی های آینده آن رشته علمی نقش بسزایی داشته باشد. در حال حاضر نزدیک به ۱۵۰۰ نشریه علمی توسط وزارت علوم، تحقیقات و فناوری اعتبارگذاری شده که بیش از ۸۵ درصد آنها دارای رتبه های الف و ب هستند که این امر نشانگر اعتبار علمی بالای آنها دارد. اما این افزایش تعداد نشریه های علمی ایجاب می کند که برای حفظ و توسعه کیفی آنها تلاش شود. از این‌رو، ارزیابی این نشریات از جهات گوناگون، می‌تواند شاخصی در جهت پایش وضعیت موجود و بهبود جایگاه نشریات کشور باشد. اما از طرفی با توجه به محدودیت‌های زمانی و همچنین حجم اطلاعات، جستجو و تحلیل دستی منابع علمی عملا امکان‌پذیر نیست و آنچه حائز اهمیت است استفاده از روش‌های یادگیری ماشین مانند متن کاوی برای پردازش داده‌های بزرگ است که در پژوهش حاضر مورد توجه قرار گرفته است. همچنین انتخاب نشریه جهت انتشار مقاله، متاثر از عوامل گوناگونی است که در این میان عامل ارتباط دامنه و حوزه

موضوعی نشریه با مقاله از اهمیت ویژه‌ای برخوردار است؛ زیرا ارسال مقالات حاصل از یک پژوهش، به نشریاتی که ربط کمتری با محتوا و نیاز خوانندگان دارند باعث اتلاف وقت پژوهشگر می‌شود. همچنین حوزه سردبیری نشریه نیز مدت زمانی را صرف بررسی و احتمالا ارسال مقاله به داوری نموده و اگر طی هر یک از مراحل مقاله رد شود این فرآیند باید مجددا برای نشریه‌ای دیگر تکرار گردد. از اینرو این نکته که نشریات تا چه اندازه در اعلام حوزه های موضوعی که مقالات را میپذیرند شفاف عمل کرده و به آن پایبند بوده اند حائز اهمیت است. گاها موضوعات اعلامی از سوی نشریات بسیار کلی بوده و کاربر را در انتخاب نشریه مرتبط با دشواری مواجه می کند.

در این راستا پژوهش حاضر با هدف تحلیل و تطبیق میزان همخوانی دامنه و اهداف موضوعی نشریات فارسی وزارت عتف با محتوای مقالات منتشر شده در انها طی بازه زمانی سه ساله ۱۳۹۷-۱۳۹۹ انجام شد تا مشخص گردد که نشریات به چه میزان در اعلام و رعایت دامنه و حوزه های موضوعی که در آن مقاله می پذیرند رویکرد شفافی داشته اند. در این پژوهش مجموعا ۱۱۶ نشریه و بالغ بر ۱۵۰۰۰ مقاله مورد بررسی و تحلیل قرار گرفت و از روش فضای برداری به منظور تعیین میزان شباهت استفاده شد. از نظر دوره انتشار، نشریات مورد مطالعه به صورت فصلنامه و دو فصلنامه منتشر می‌شوند. همچنین نشریات مورد مطالعه از نظر پوشش موضوعی به ترتیب از حوزه‌های علوم انسانی، فنی و مهندسی، علوم پایه، هنر و معماری، کشاورزی و منابع طبیعی و دامپزشکی می باشند.

یافته‌های پژوهش در میزان رعایت و پایبندی نشریات به اهداف و دامنه موضوعی اعلام شده و محتوای مقالات منتشر در هر یک از انها نشان داد تنها در ۸ درصد نشریات مورد بررسی به میزان ۹۵ درصد تطابق وجود داشت و نویسندگان را به طور شفاف از موضوعات مورد پذیرش نشریه آگاه کرده و خود نیز پایبند به حوزه های موضوعی اعلام شده بوده اند. در ۱۹ درصد نشریات اعلام موضوعات بسیار کلی بوده و تطابق نیز تنها در حد ۴۰ درصد می باشد.

در مطالعات پیشین نیز پژوهشگران به این نتیجه رسیده اند که درصدی از مقالات منتشر شده در نشریات با موضوع های اعلامی همخوانی ندارند از جمله نتایج مطالعه افضلی پور و جمالی مهموئی در بررسی میزان تخصص گرایی مجله های حوزه علم اطلاعات و دانش شناسی فارسی طی سالهای ۱۳۸۷ الی ۱۳۹۱ نشان داد ۷ درصد مقالات با موضوعای غیرمرتبط با

علم اطلاعات در این نشریات منتشر شده اند. همچنین تحلیل محتوایی مقالات نشریه اخلاق در علوم و فناوری نشان داد موضوع هایی از جمله کدهای اخلاقی، رهبری اخلاقی، تعهد سازمانی و .. که بیشترین تعداد مقالات را طی سالهای ۸۵ تا ۹۴ به خود اختصاص داده اند در اداف و دامنه موضوعی نشریه ذکر نشده اند (غلامی، ۱۳۹۵). بیشترین سهم موضوعی مقالات در نشریه آموزش و توسعه منابع انسانی طی سالهای ۹۳ تا ۹۷ مربوط به موضوع اندازه‌گیری اثرات و نتایج برنامه های آموزش و توسعه و سپس مباحثی از قبیل سنجش اثربخشی دوره های آموشی، آموزش و یادگیری الکترونیکی، مدیریت دانایی بود (روحانی راد و همکاران، ۱۳۹۹) که در اهداف و دامنه موضوعی نشریه هیچ اشاره ای بدانها نشده است.

هرچند در برخی نشریات مانند نشریه مطالعات مدیریت گردشگری، گرایشهای موضوعی مقالات طی سالهای ۱۳۹۱ تا ۱۳۹۸ در نشریه مطالعات مدیریت گردشگری نشان داد مدیریت توریسم، تبلیغات توریسم، اقتصاد توریسم و مدیریت هتلها به ترتیب دارای بیشتریت تعداد مقاله هستند (رضانیا و همکاران، ۱۴۰۰) و با دامنه و حوزه موضوعی مدرج در وبگاه نشریه مطابقت دارد. همچنین در نشریه علمی راهبرد با توجه به یافته های پژوهش آخوندی (۱۳۹۸) همخوانی لازم به میزان مطلوبی رعایت شده است. بر اساس یافته های مطالعه خاشعی و همکاران (۱۳۹۸) و مطابق یا مقالات منتشر شده در این نشریه طی سالهای ۹۰ تا ۹۶، نشریه مطالعات مدیریت بهبود و تحول نیز به صورت شفاف موضوعات مورد پذیرش نشریه را در وبگاه اطلاع رسانی نموده است.

در پیوند با مطالب پیش گفته می توان اینچنین مطرح کرد که در پیش گرفتن رویکرد شفاف موضوعی از سوی نشریات می تواند دستاوردهای سودمندی برای جامعهٔ دانشگاهی داشته باشد. با مشخص بودن شدن گرایش های موضوعی نشریه های علمی، هویت بارزتری برای هر نشریه ایجاد می شود؛ به طوری که هر متخصصی می‌تواند انتظار داشته باشد که نشریه یا نشریه های خاصی، زمینه های مطالعاتی وی را پوشش خواهد داد. تخصص گرایی فرآیندی علمی و عملی است که می توان به یاری آن، پژوهش های اعضای هیات علمی و پژوهشگران کشور را معطوف به حوزه های تخصصی کرد. در تعریف حوزه تخصص باید گفت که حوزه تخصص به آن حوزه یا گرایش خاصی اطلاق می گردد که مشتمل بر مجموعه مسائل نظام مندی است که نقشه راه فعالیت نشریه را در بازه زمانی طولانی تعیین می نماید و نتیجه آن به فراوری دانش انباشته و رسیدن به

نظریه های جدید در آن حوزه علمی منتهی می گردد. بدین ترتیب، نشریه و نویسندگان آن به مرجعی در جامعه علمی تبدیل می گردند.

۵- ۳ پیشنهادهای کاربردی

پیشنهاد می‌شودگروه دبیران، ضمن بررسی منظم روند انتشار مقالات در مجلات خود، از سیر موضوعی مقالات همراستا با پژوهش‌های جهانی اطمینان حاصل نمایند و ضمن انعطاف پذیری در موضوعات مورد پذیرش، از انحراف مجله تحت مدیریت خود از خط مشیء آن جلوگیری کنند. همچنین با توجه به گرایش های موضوعی در گذر زمان، می‌توانند از همپوشانی موضوعی مجلات ممانعت به عمل آورده و موضوعات مغفول را در دامنه موضوعی مجلات بگنجانند. در این راستا به منظور تعدیل وضعیت موجود پیشنهاد میشود ضمن استخراج بیشترین واژه های کلیدی به کار رفته شده در مقالات، فهرست حوزه‌های موضوعی مندرج در وبگاه نشریه را روزآمد نمایند. این امر علاوه بر اینکه از ارسال مقالات غیر مرتبط با موضوع نشریه و اتلاف زمان و هزینه پژوهشگر و دست اندرکارن نشریه جلوگیری می‌شود، تخصصی شدن و عدم پراکندگی موضوعی نشریه را نیز به دنبال خواهد داشت. همچنین پیشنهاد می شود پژوهشگران نیز پیش از انجام پژوهش، ضمن بررسی موضوعات روز، به مجلاتی که در زمینه های موضوعی مرتبط و تخصصی اقدام به انتشار مقاله می کنند، توجه خاص تری داشته باشند.

۵-۴. پیشنهادهای پژوهش

- پیشنهاد می‌شود پژوهشی مشابه در خصوص مجلات ایرانی بین‌المللی انجام و نتایج پژوهش حاضر با پژوهش پیشنهادی مقایسه گردد.
- پیشنهاد می شود پژوهش باابررسی دیگر نشریات مجددا تکرار و در صورت کسب نتایج مشابه، شفاف سازی ذکر دامنه موضوعی نشریات از سوی کمیسیون به انها ابلاغ گردد

فهرست منابع

- آخوندی، عباس. (۱۳۹۸). تحلیل محتوای مقالات فصلنامه راهبرد. *فصلنامه علمی راهبرد، ۲۸(۲): ۱۶۷-۱۹۶*. dor: 20.1001.1.10283102.1398.28.2.7.0
- استادزاده، زیبا؛ داوودی، حسین؛ حیدری، حسن؛ میرمهدی، سیدرضا (۱۳۹۷). تحلیل محتوای مقالات «دوفصلنامه مطالعات اسلام و روان‌شناسی(۱۳۹۵-۱۳۸۶). *مطالعات اسلام و روانشناسی*. ۱۲ (۲۳): ۱۴۱-۱۵۸.
- افضلی پور حدیثه، جمالی مهموئی حمید رضا (۱۳۹۳). میزان تخصص گرایی مجله‌های علمی- پژوهشی علم اطلاعات و دانش‌شناسی فارسی بر اساس موضوع مقالات منتشر شده بین سال‌های ۱۳۸۷ تا ۱۳۹۱. تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی. ۲۰ (۳) :۴۲۲-۳۹۷.
- اللهیاری، محمدصادق؛ دقیقی ماسوله، زهرا؛ افتخاری، هاجر (۱۳۹۴). تحلیل محتوای مقالات تخصصی علوم ترویج و آموزش کشاورزی در ایران (۱۳۹۲- ۱۳۸۸). علوم ترویج و آموزش کشاورزی، ۱۱(۱)، ۲۲۹-۲۴۹.
- باغ محمد، مریم، منصوری، علی، چشمه سهرابی، مهرداد. (۱۴۰۱). بررسی توسعه و روند موضوعی حوزه علم اطلاعات و دانش‌شناسی بر اساس مدل موضوعی LDA. *پژوهشنامه پردازش و مدیریت اطلاعات، ۳۶* (۲): ۲۹۷-۳۲۸ . doi: 10.35050/JIPM010.2020.001
- توکلی زاده راوری، محمد؛ دهقانی، نجابتیان؛ سهیلی. (۱۳۹۴). تحلیل محتوای مقالات فارسی نشریات علمی ایران در زمینه ازدواج و طلاق با روش خوشه‌بندی سلسله مراتبی. *فصلنامه فرهنگی تربیتی زنان و خانواده، ۱۰* (۳۲).
- جانانی، پیمان؛ رودباری؛ تهمتن؛ صدقی (۱۳۹۱). تحلیل محتوای مقالات نشریات دانشکده‌های پرستاری و مامائی دانشگاه‌های علوم پزشکی ایران. مراقبتهای پرستاری و مامایی،. ۲ (۱)، ۵۳-۶۱.
- حاتمی ناغانی, بهمنو عابسی, مسعود. (۱۳۹۵). تحلیل محتوایی مقالات علمی با استفاده از متن‌کاوی .*مطالعات مدیریت کسب و کار هوشمند* 10.22054/ims.2017.7014 .doi: 137-167 ,(18)5 ,
- حری، عباس . (۱۳۷۴). بررسی رابطه میان مجلات منتشره حوزه های تخصصی وارتقاء علمی متخصصان کشور.

پژوهش‌های کاربردی روانشناختی ۱ (۲): ۲۳-۴۴.

- حسینی نسب، اعظم؛ محمدی، مهدی؛ طالعی، عبدالحسین (۱۳۹۵). تحلیل موضوعی مقالات علمی پژوهشی حوزه علوم قرآن؛ نشر یافته بین سال‌های ۱۳۸۶-۱۳۹۰. دو فصلنامه کتاب قیم. ۶ (۱۵): ۱۴۱-۱۶۸.
- خاشعی ورنامخواستی، وحید؛ طیبی ابوالحسنی، سیدامیرحسین؛ اسدی خانقاه، شیرین (۱۳۹۸). تحلیل محتوای مقالات فصلنامه مطالعات مدیریت (بهبود و تحول) طی دوره ۷ ساله. (۱۳۹۰-۹۶). مطالعات مدیریت (بهبود و تحول): ۲۸ (۹۳): ۱۲۹-۱۶۰.
- داستانی، میثم؛ موسوی چلک، افشین؛ ضیائی، ثریا؛ دلقندی، فائزه. (۱۳۹۹). تجزیه و تحلیل موضوعی مقالات منتشرشده ی کتابداری و اطلاع رسانی پزشکی در ایران با استفاده از فنون متن کاوی. تصویر سلامت ۱۱ (۴): ۳۵۵-۳۶۷.doi: 10.34172/doh.2020.43
- دهدشتی شاهرخ، زهره. (۱۳۹۹). تحلیل ساختار محتوایی فصلنامه مطالعات مدیریت گردشگری با استفاده از تکنیک متن‌کاوی. *مطالعات مدیریت گردشگری ۱۵ (۱۵): ۹۷-۱۲۷*. doi: 10.22054/tms.2020.42515.2141
- رضانیا، درنا، مقنی پور، مجید رضا و ظفرمند، سید جواد. (۱۴۰۰). تحلیل محتوای مقالات فصلنامهٔ مطالعات مدیریت گردشگری (۱۳۹۱-۱۳۹۸) برمبنای الگوی مشارکت نویسندگان و ویژگی‌های روش‌شناسی، استنادی و عملکردی مقالات .*فصلنامه بازیابی دانش و نظام‌های معنایی* doi: 169-201. ,*8*(28), 10.22054/jks.2021.60526.1433
- سلک، محسن، بزرگی، اشرف السادات. (۱۳۸۹). تحلیل محتوای مقالات منتشر شده در دو نشریه فصلنامه "کتابداری و اطلاع رسانی" و " فصلنامه کتاب" در سالهای ۱۳۸۵و۱۳۸۶. دانش شناسی. ۳ (۱۰): ۲۵-۴۰.
- شکوهیان، محبوبه؛ عاصمی، عاصفه؛ شعبانی، احمد؛ چشمه سهرابی، مظفر. (۱۴۰۱). ارائه مدل دسته‌بندی موضوعی تولیدات علمی حوزه سلامت با استفاده از روش‌های متن‌کاوی. *پژوهشنامه پردازش و مدیریت اطلاعات ۳۵ (۲): ۵۵۳-۵۷۴*. doi: 10.35050/JIPM010.2020.061 .

- صفوی جهرمی، گلایل، طباطبائیان، سید حبیب الله، حنفی زاده، پیام، حاجی میرزائی، حامد. (۱۳۹۹). نقشه موضوعی مقالات حوزه تولید محتوای دیجیتال برای کودکان و نوجوانان. مطالعات کتابداری و علم اطلاعات. ۱۴ (۳).
- غلامی، طاهره. (۱۳۹۵). تحلیل محتوای مقالات فصلنامه اخلاق در علوم و فنّاوری (۱۳۸۵ – ۱۳۹۴). اخلاق در علوم و فناوری. ۱۱ (۱) :۳۸-۲۹
- فتاحی، رحمت الله. (۱۳۹۴) تخصصی کردن گرایش موضوعی مجله های علمی و چالشهای آن. پژوهشنامه پردازش و مدیریت اطلاعات. ۳۰ (۳): ۶۰۱-۶۰۲.
- قناد، مصطفی؛ عرب مازار یزدی، محمد؛ صفرزاده بندری؛ محمدحسین؛ حصارزاده، رضا. (۱۴۰۲). کاربست فنون متن‌کاوی در تحلیل جریان موضوعی مقالات منتشره در مجلات حسابداری ایران. *پژوهش های حسابداری مالی و حسابرسی ۱۵(۵۸): ۱-۳۸.*
- کاظمی، عبدالحسن. (۱۳۸۸) مشکلات سنجش تولیدات علمی کشور. مدیریت اطلاعات سلامت. فروردین ۱۳۸۸.
- Abel, R.E. & Newlin, L. W. (2002). *Scholarly Publishing: Books Journals, Publishers, and Libraries in the Twentieth Century*. N.Y.: ISBN 0-471-21929-0.Y.: Wiley.
- Abramson, D., Lees, M., Krzhizhanovskaya, V. V., Dongarra, J. J., & Sloot, P. M. (2014). Big Data Meets Computational Science, Preface for ICCS. In *ICCS.* 1-7.
- Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6*(1):147-153.
- Alwidian, S. A. A., Bani-Salameh, H. A., Alslaity, A. A. N. (2015). Text data mining: a proposed framework and future perspectives. *International Journal of Business Information Systems*, *18*(2): 127-140.
- Anderson, K. (2012). Editorial Rejection — Increasingly Important, Yet Often Overlooked Or Dismissed, in The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2012/04/19/post-publication-peerreview-what-value-do-usage-based-metrics-offer.

- Ananiadou, S., McNaught, J. (2006). *Text mining for biology and biomedicine*. London: Artech House.
- Baars, H., Kemper, H. G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management*, *25*(2): 132-148.
- Blei DM. (2012). Probabilistic topic models. *Communications of the ACM*. 55(4): 77-84.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9, 13.
- Chen, H., Huang, X., & Li, Z. (2022). A content analysis of Chinese news coverage on COVID-19 and tourism. *Current Issues in Tourism*, *25*(2), 198-205.
- Choudhary AK, Oluikpe PI, Harding JA, Carrillo PM. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. Computers in Industry. 1;60(9):728-40.
- Clifton, C., Cooley, R., & Rennie, J. (2004). Topcat: Data mining for topic identification in a text corpus. *IEEE transactions on knowledge and data engineering*, 16(8): 949-964.
- Davies، K. (2012). Content analysis of research articles in information systems (LIS) journals. *Library and Information Research*، *36*(112)، 16-28.
- Delen, D. (2014). *Real-world data mining: applied business analytics and decision making*. NewJersey: Financial Times Press
- Dong, G., Liu, H. (Eds.). (2018). *Feature engineering for machine learning and data analytics*. CRC Press.
- Elo S, Kyngäs H. (2008). The qualitative content analysis process. Journal of advanced nursing. 62(1):107-15.
- Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information processing & management*, 41(6): 1548-1572.

- Golub, K. (2006). Using controlled vocabularies in automated subject classification of textual web pages, in the context of browsing. *IEEE TCDL Bulletin*, 2(2): 1-11.
- Hofmann T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*; Berkeley, California, USA: Association for Computing Machinery; 50–57.
- Huisman, J., and J. Smits. (2017). Duration and quality of the peer review process: the author's perspective. Scientometrics, 113 (1): 633-650.
- Ilias, K. (2019). A distributed text claustering approach for detecting duplicate questions (Doctoral dissertation, Aristotle University of Thessaloniki Thessaloniki, Greece).
- Indurkhya, N., Damerau, F. J. (Eds.). (2010). Handbook of natural language processing (Vol. 2). CRC Press.
- Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology*, 2(11): 2319-1163.
- Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information processing & management*, 42(6): 1614-1642.
- Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
- Khan, R. A. (2016). Data Mining: A Tool for Customer Relationship Management. *Data Mining and Knowledge Engineering*, *8*(4), 95-99.
- Kim, Y., M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. Health informatics journal, 24(4): 432-452.
- Lamba, M., & Madhusudhan, M. (2019). Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. Scientometrics, 120(2): 477-505
- Lin, Z., S. Hou, and J. Wu. (2016). The correlation between editorial delay and the ratio of highly cited papers in Nature, Science and Physical Review Letters.

Scientometrics 107 (3): 1457-1464.

- Majhi, S., Jal, Ch. & Maharana, B. (2016). Content analysis of Journal articles on Wiki in
- Science Direct Database. *Library Philosophy and Practice (e-journal),* 1 – 15.
- Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information systems*, 34(1): 87-107.
- Marinakos, G., Daskalaki, S. (2016). Viability prediction for retail business units using data mining techniques: a practical application in the Greek pharmaceutical sector. *International Journal of Computational Economics and Econometrics*, 6(1): 1-12.
- Mesropyan, V. R., Ovsyannikov, M. V. (2014). Prospects for the application of scientometric methods for forecasting. *Scientific and Technical Information Processing*, 41(1): 38-46.
- Mitra, S., Acharya, T. (2003). *Data Mining: Concepts and Algorithms From Multimedia to Bioinformatics*. John Wiley & Sons.
- Mulligan, A., Hall, L. & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. Journal of the American Society for Information Science and Technology, 64(1), 132-161.
- Nguyen, T. T., F. Maxwell Harper, L. Terveen, et al. (2018). User Personality and User Satisfaction with Recommender Systems. Information Systems Frontiers 20 (6): 1173-1189.
- Nicholson, S., Hwang, S. Y., Keezer, P., & O'Neill, E. T. (2003). The bibliomining process: Data warehousing and data mining for libraries. Sponsored by SIG LT. *Proceedings of the American Society for Information Science and Technology*, 40(1): 478-479.
- Pranata, I., Skinner, G. (2015). Segmenting and targeting customers through clusters selection & analysis. In *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS),* 303-308.

- Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning. 242 (1): 29-48.
- Saheb T, Saheb M. (2019). Analyzing and visualizing knowledge structures of health informatics from1974 to 2018: A bibliometric and social network analysis. *Healthc Inform Res.* 25(2): 61-72.
- Salloum SA, Al-Emran M, Monem AA, Shaalan K. (2017). A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. 2(1):127-33.
- Sharma, N. (2005). *Discovering knowledge with text mining*. A&M University-Kingsville.
- Silva, R. O., de Carvalho, R. L. (2021). Using Latent Semantic Indexing as a metric for evaluating research potentialities through Innovation Public Policies. Academic Journal on Computing, Engineering and Applied Mathematics, 2(2): 10-15.
- Spinakis, A., Peristera, P. (2004). Text mining tools: Evaluation methods and criteria. In *Text Mining and its Applications* (pp. 131-149). Springer, Berlin, Heidelberg.
- Srivastava AN, Sahami M. (2009). Text Mining: Classification, Clustering, and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) 1st Ed. Uk: Chapman and Hall/CRC.
- Smith SD, Ng K-M, Brinson J, Mityagin E. (2008). Multiculturalism, diversity, and social advocacy: A 17-year content analysis of counselor education and supervision. Counselor Education and Supervision.47(4):249.
- Soleimani Nezhad A, Salajegheh M, Tayyebi Nia E. (2019). Clustering scientific articles based on the k_means algorithmCase Study: Iranian Research Institute for information Science and Technology . Iranian Journal of Information Processing and Management. 34(2):871-96.
- Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes*, *26*(1), 90–102.

- Stroud، D.، Pennington، P.، Cleaver، C.، Collins، J. R.، & Terry، N. (2017). A content analysis of research articles in The Journal for Specialists in Group Work: 1998–2015. *The Journal for Specialists in Group Work*، *42*(2)، 194-210.
- Wang C, Blei D, Heckerman D. (2008). Continuous time dynamic topic models. *Proceedings of the Twenty- Fourth Conference on Uncertainty in Artificial Intelligence*; Helsinki, Finland: AUAI Press; 579–86.
- Wang, W. T., and Y. P. Hou. (2015). Motivations of employees' knowledge sharing behaviors: A selfdetermination perspective. Information and Organization 25 (1): 1-26.
- Wei X, Croft WB. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*; Seattle, Washington, USA: Association for Computing Machinery; 178–85.
- Witten, I. H., Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, 31(1): 76-77.
- Woo H, Heo N. (2013). A content analysis of qualitative research in select ACA journals (2005–2010). Counseling Outcome Research and Evaluation. 4(1):13-25.
- Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1): 118-133.
- Yau C-K, Porter A, Newman N, Suominen A. (2014). Clustering scientific documents with topic modeling. *Scientometrics.* 100(3): 767-86.
- Yao, W., J. He, H. Wang, Y. Zhang, & J. Cao. (2015). Collaborative topic ranking: Leveraging item metadata for sparsity reduction. In Proceedings of the AAAI Conference on Artificial Intelligence 29 (1): 374-380.
- Zeigler EF. (1987). Sport Management: Past, present, future. Journal of sport management. 1(1):4-24.