
بسم الله
الرحمن
الرحيم

وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
پژوهشکدهٔ فناوری اطلاعات

رسالهٔ دکترا

رشتهٔ مهندسی فناوری اطلاعات

تحلیل‌روند علمی کشور و پیش‌بینی فناوری با استفاده از روش‌های یادگیری ماشین. مورد مطالعه: سامانه گنج

نویسنده

اشکان خطیر

استاد راهنما

دکتر سهیل گنجه‌فر

استاد مشاور

دکتر آزاده محبی

بهمن ۱۳۹۷

اصالت و مالکیت رساله

این‌جانب اشکان خطیر دانش‌آموختهٔ دکتری رشتهٔ مهندسی فناوری اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) پدیدآور رساله با عنوان تحلیل‌روند علمی کشور و پیش‌بینی فناوری با استفاده از روش‌های یادگیری ماشین. مورد مطالعه: سامانه گنج با راهنمایی دکتر سهیل گنجه‌فر گواهی و تعهد می‌کنم که بر پایهٔ قوانین و مقررات، از جمله «دستورالعمل نحوهٔ بررسی تخلفات پژوهشی» و همچنین «مصادیق تخلفات پژوهشی» مصوب وزارت علوم، تحقیقات و فناوری (۲۵ اسفند ۱۳۹۳):

- این رساله دستاورد پژوهش این‌جانب و محتوای آن از درستی و اصالت برخوردار است؛
- حقوق معنوی همهٔ کسانی را که در به‌دست‌آمدن نتایج اصلی رساله تأثیرگذار بوده‌اند، رعایت کرده‌ام و هنگام کاربرد دستاورد پژوهش‌های دیگران در آن، با دقت و به‌درستی به آن‌ها استناد کرده‌ام؛
- این رساله و محتوای آن را تاکنون این‌جانب یا کس دیگری برای دریافت هیچ‌گونه مدرک هم‌سطح یا بالاتر در هیچ‌جا ارائه نکرده‌ایم؛
- همهٔ حقوق مادی این رساله از آن پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است و آثار اصلی برگرفته از آن با وابستگی سازمانی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) منتشر خواهند شد؛
- در همهٔ آثار اصلی برگرفته از این رساله، نام استاد(ان) راهنما و اگر استاد راهنمای نخست تشخیص دهد، نام استاد(ان) مشاور و نشانی رایانامهٔ سازمانی آنان را می‌آورم؛
- در همهٔ گام‌های انجام این رساله، هرگاه به اطلاعات شخصی افراد یا اطلاعات سازمان‌ها دسترسی داشته یا آن‌ها را به‌کار برده‌ام، رازداری و اخلاق پژوهش را رعایت کرده‌ام.

تاریخ **امضا**

تقدیم به

خانواده عزیزم.

سپاس

با سپاس از اساتید عزیز و محترمم جناب آقای دکتر گنجه‌فر و سرکار خانم دکتر محبی که درس‌هایی فراتر از رساله جاری به من آموختند.

با سپاس از خانواده پژوهشگاه علوم و فناوری اطلاعات و رئیس پژوهشگاه .

چکیده

هدف: شرکت‌ها و سازمان‌ها برای بقا و رقابت نیازمند بررسی روند فعالیت خود و رقبا هستند. وجود یک مسیر و نقشه از فعالیت‌های موجود از جمله نیازهای اولیه برای بقا است. بخاطر خلاء سیستمی که به صورت خودکار و به دور از تعصب فکری بتواند روند حوزه‌های علمی را مشخص کند و ارتباط بین موضوع‌های علمی را در زبان فارسی نشان دهد باعث شد تا در این رساله هدف روشی برای تحلیل‌روند و پیش‌بینی فناوری را با استفاده از یادگیر ماشین به صورت کاملا خودکار ارائه کنیم. بنابراین هدف از این پژوهش ارائه قالب کاری و سامانه‌ای برای تحلیل‌روند و رشد علم و فناوری بر اساس پایان‌نامه‌ها و رساله‌های فارسی ثبت شده دانشگاه‌ها در سامانه گنج است.

روش پژوهش: تحلیل‌روند خودکار در چهار گام اصلی انجام شده است: بررسی و شناخت داده‌ها، تحلیل‌خودکار متون و استخراج خودکار کلیدواژه‌ها، بازشناسی الگو و مدل‌سازی موضوعی و تحلیل و ارزیابی روند که هرگام چالش‌های خود را دارد. بعد از مطالعه روش‌های تحلیل‌روند علمی و فناوری و استخراج نیازمندی‌ها به شناخت خصوصیات اسناد در دسترس پرداخته می‌شود (که در این رساله پارسا ثبت شده در سامانه گنج هستند). برای این منظور اسناد مورد نیاز استخراج و ساختار چکیده‌ها و کلیدواژه‌های پارسا را مورد بررسی و ارزیابی قرار خواهیم داد. سپس اسناد را باید به شکل مناسبی قابل پردازش برای سیستم‌های دیجیتال تبدیل کنیم. بنابراین با استفاده از نتایج شناخت خصوصیات اسناد کلیدواژه‌هایی به صورت خودکار استخراج می‌شود و با استفاده از روش‌های بازنمایی کلمات، مجموعه‌داده‌های هم‌رخدادی کلمات را تشکیل خواهیم داد. در گام بعدی بر اساس اسناد تحلیل‌شده و الگوهای رفتاری بین کلمات استخراج شده، موضوع‌های مخفی موجود در متون را استخراج خواهیم کرد. در ادامه شاخص‌های تحلیل‌روند را برای موضوع‌های استخراج شده محاسبه کرده و با استفاده از نمودار راهبردی موضوعات را بصری‌سازی و روند رشد آن‌ها مورد بررسی و تحلیل قرار خواهیم داد.

یافته‌ها و نتایج: در این رساله علاوه بر ارائه روشی برای اولین بار برای تحلیل‌روند خودکار متون فارسی برای هر گام از پژوهش ریکردهای جدیدی ارائه کردیم و نتایج حاصله را مورد ارزیابی قرار داده‌ایم. در گام نخست با بررسی اسناد فارسی نحوه توزیع کلیدواژه‌ها در عنوان و چکیده مشخص شد و نشان داد شده که ۵۵ ٪ از نمایه‌ها استانداردهای چکیده‌نویسی را رعایت نکرده‌اند و ۳۸ ٪ از آن‌ها تعداد کلیدواژه‌های کافی برای اسناد خود انتخاب نکرده‌اند. توزیع استخراج‌شده نشان می‌دهد که برای استخراج کلیدواژه برای پارسا، استفاده از عنوان و نمایه‌های موجود می‌تواند کافی باشد. روش ارائه‌شده برای انتخاب کلیدواژه‌ها در گام بعدی با کاهش ۷۸ درصدی کلیدواژه‌ها تنها حدود ۳ درصد افت در صحت دسته‌بندی داشته است. مقایسه‌ی دیگری که در این رساله بین دسته‌بندهای مختلف برای اسناد علمی فارسی انجام شد نشان داد که دسته‌بند درخت بیشترین میزان صحت (در حدود ۹۹ ٪ درصد) را در بین دسته‌بندهای

بیزین، فرا دسته‌بندها (بگینگ) و مدل‌های رگرسیونی داشته است. با آزمودن خوشه‌بندهای مختلف و محاسبه خلوص برای روش ارائه‌شده در بازنمایی کلمات در این رساله در مقایسه با روش‌های بازنمایی مختلف و مقایسه دسته‌بندهای مختلف نشان داده‌شده که روش ارائه‌شده با اختلاف ۲۰ درصدی (میانگین ۷۵ ٪) در میزان خلوص از مابقی روش‌های خوشه‌بندی بهتر عمل کرده است. در مقایسه خوشه‌بندهای مختلف، استفاده از روش کا-میانه و خوشه‌بند فازی با m=۲.۱ بیشترین میزان خلوص در خوشه‌بندی را ارائه داده‌اند. افزایش نتایج خوشه‌بندی نشان دادند که استفاده از موضوع‌ها بجای کلمات می‌توانند باعث حذف کلمات نویز در پارسا شوند و تحلیل‌های دقیق‌تری را نشان دهند که بدین منظور نمودار استراتژیک برحسب موضوع ارائه گردید.

جمع‌بندی: در این رساله برای اولین بار روشی خودکار برای تحلیل‌روند و استخراج موضوع‌های علمی فارسی ارائه‌شده است. در ساختار آن نیز روش‌هایی برای بازنمایی کلمات، انتخاب کلیدواژه و تحلیل‌روند با استفاده از نمودار راهبردی موضوعی ارائه گردید. مقا یسه ها نشان‌دهنده بهبود در هر یک از روش‌های پیشنهادی است.

کلیدواژه‌ها: تحلیل‌روند، یادگیری ماشین، نمودار راهبردی، مدل‌سازی موضوعی، بازنمایی کلمات، خوشه‌بندی متون، استخراج ویژگی، دسته‌بندی خودکار متون علمی

فهرست

فهرست جدول‌ها

فهرست شکل‌ها

فصل ۱- کلیات تحقیق

۱-۱- مقدمه

امروزه بخاطره پویایی و سرعت رشد تولیدات علمی، تغییر و تحولات فناوری سرعت زیادی به خود گرفته است. شرکت‌ها و سازمان‌های برای رقابت‌پذیری نیازمند برنامه‌ریزی دقیق و بررسی روند پیشرفت خود در طول این برنامه‌ریزی هستند. بنابراین ابزارهای تحلیل‌روند و پیش‌بینی فناوری را می‌توان از ملزومات یک سیستم مدیریتی کارامد در نظر گرفت. تحلیل‌روند یکی از روش‌های پیش‌بینی است. بنابراین با تحلیل‌روند یک فناوری می‌توان آن فناوری را پیش‌بینی کرد. از طرفی«فناوری» را می‌توان: ابزارها، شیوه‌ها و فرایندهایی که برای برآورده کردن نیازهای انسان مورداستفاده قرار می‌گیرد (Martino۱۹۹۳) تعریف کرد. لذا پیش‌بینی فناوری به‌طورکلی هرگونه روش هدفمند و نظام‌مندی برای تحلیل، درک و پیش‌بینی یک فناروری گفته می‌شود که آن فناوری را در زمینه‌های نرخ رشد، جهت رشد و تغییرات مورد بررسی قرار خواهند گرفت (بخصوص در زمینه ابداعات، نوآوری‌ها) (Firat, Woon, and Madnick۲۰۰۸). از سویی دیگر تعاریف دیگری در خصوص پیش‌بینی فناوری ارائه شده است که عبارت است از: پیش‌بینی تمایلات و روند فناوری مانند جهت رشد و نرخ توسعه فناوری (Jun and Lee ۲۰۱۲, Jun ۲۰۱۱, Porter ۱۹۹۱a).

بنابراین می‌توان نتیجه گرفت با تحلیل و بررسی روند روش‌ها، شیوه‌ها و فرایندهای در حال‌ظهور می‌توان فناوری‌های بعدی را تا حدودی پیش‌بینی کرد. از جمله اسنادی که ابزارها، روش‌ها و شیوه‌های جدید معرفی می‌کنند پژوهش‌ها و فعالیت‌های انجام شده در تحصیلات دانشگاه‌ها است که خروجی آن‌ها عموما به صورت رساله‌ها و پایان‌نامه‌های و مقالات است. بنابراین با استفاده از تحلیل و بررسی این اسناد علاوه بر تحلیل‌روند گذشته حوزه‌های علمی قادر خواهیم بود تا حدودی از ابزارها، روش‌ها و فناوری‌هایی که در آینده قابلیت عملیاتی شدن را دارند را پیش‌بینی کنیم.

۱-۲- بیان مسئله

هر سازمان و شرکتی نیاز به برنامه‌ریزی دارد. این برنامه‌ریزی با دانستن روند حرکت فعلی و وضعیت آینده می‌تواند به نحو بهتری صورت پذیرد. بنابراین دانش وضعیت فعلی و پیش‌بینی آینده ابزارهای قدرتمندی هستند که برای برنامه‌ریزی می‌توانند مورداستفاده قرار گیرند. پیش‌بینی آینده فناوری برای تمام درجات سازمانی کشور چه از سازمان‌های بزرگ و وزارتخانه‌ها و چه برای بنگاه‌های کوچک بسیار بااهمیت است. پیش‌بینی فناوری در تعیین سیاست‌ها و چشم‌اندازهای علمی، تحقیقاتی و فناوری نقش بسیاری دارد. بنابراین انتخاب یک روش قابل استناد که عاری از هرگونه تمایلات شخصی و پیش‌فرض‌های علمی باشد، بسیار مهم است.

به‌طورکلی روش‌های پیش‌بینی فناوری را به پنج قسمت تقسیم کرده‌اند، استفاده از نظرات متخصصان، تحلیل‌روند، پایش اطلاعات، مدل‌سازی و سناریونویسی (Porter۱۹۹۱). پایش یک فرایندی است که نشانه‌های یک فناوری را در یک حوزه خاص شناسایی می‌کند و امکان آگاهی از نشانه‌هایی که نفوذ و پذیرش آن فناوری را بیان می‌کند به ما نشان می‌دهد. بر اساس آن نشانه‌ها ما قابلیت پیش‌بینی فناوری را خواهیم داشت. نظرات متخصصان بر این اصل استوار است که افراد دارای تخصص درزمینه‌ی تخصصی خود قابلیت پیش‌بینی بهتری دارند. در تحلیل‌روند، روش‌های آماری روند رشد فناوری را موردبررسی قرار می‌دهند و آن را به آینده بسط خواهند داد. با این روش آینده را می‌توان تا حدودی مورد پیش‌بینی قرارداد. مدل‌سازی می‌تواند رفتار آینده یک سیستم پیچیده را به‌سادگی نشان دهد. البته درست کردن یک مدل بسیار سخت و لازمه محاسبات بسیار پیچیده و در نظر گرفتن تمام شرایط ممکن است. در سناریونویسی نیز ابتدا به تشخیص رویدادها و یا رخدادهای کلیدی که ممکن است در تکامل آینده فناوری نقش داشته باشند، می‌پردازید. سپس بر اساس آن‌یک سری تصاویر ترسیمی از آینده را ترسیم می‌کند و آینده پیش‌بینی می‌شود.

بعضی از این روش‌ها بر اساس قضاوت و نظر متخصصان صورت می‌گیرد. داشتن پیش‌فرض‌ها، تعصبات و تمایلات شخصی در استفاده از نظرات متخصصان می‌تواند باعث اشتباه در آینده‌نگری شود و برنامه‌ریزی‌های بلندمدت و میان‌مدت را به خطر اندازد. روش سناریونویسی برای مواردی مورداستفاده قرار می‌گیرد که اطلاعاتی از دوره‌های زمانی گذشته در رابطه با گذشته یک فناوری در دسترس نباشد یا متخصصان درزمینه‌ی مربوط ضعیف بوده یا وجود نداشته باشند. هر دو روش‌های

فوق روش‌های مبتنی بر انسان هستند و بنابراین زمان زیادی برای پیش‌بینی صرف خواهد شد. در مدل‌سازی علاوه بر پیچیده بودن و زمان‌بر بودن ساخت یک مدل برای فناوری یک کشور، این احتمال وجود دارد که تمام حالات ممکن در نظر گرفته نشده باشند و مدل درجاهای حساس به‌درستی عمل نکند. در ضمن در مدل‌سازی داده‌های تاریخی آن فناوری مدنظر قرار نمی‌گیرند.

استفاده از داده‌های عددی باعث کم‌تر شدن پیش قضاوت‌ها و تعصبات نظرات اشخاص خواهد شد و علاوه بر افزایش دقت در پیش‌بینی باعث قابلیت اسناد و مکتوب شدن پیش‌بینی خواهد شد. روش‌های آماری این ویژگی‌ها را دارند. اما درزمانی که از این روش‌ها استفاده می‌کنیم امکان دارد شرایطی برقرار باشد که در زمان گذشته وجود نداشته باشد و می‌تواند در پیش‌بینی آینده ما خدشه ایجاد کند. ازآنجاکه برون‌یابی روند از داده‌های تاریخی استفاده می‌کند، این مشکل را در خود دارد. از طرفی دیگر داده‌ها باید به‌صورت کاملاً عددی در اختیار این سیستم گذاشته شوند و قابلیت استفاده از اطلاعاتی که به‌صورت نهفته در اسناد و مدارک وجود دارند را نمی‌توانند در نظر بگیرند. از طرفی دیگر کمبود در داده‌های عددی و روش محاسباتی اشتباه را می‌توان از معایب این شیوه دانست. ازآنجاکه پیش‌بینی غلط می‌تواند زیان‌بارتر از پیش‌بینی نکردن باشد، نیاز به یک سیستمی که بتواند عمل پیش‌بینی را با دقت بالایی انجام دهد، احساس می‌شود.

از آنجا که روشی خودکار برای تحلیل‌روند علم و فناوری بر اساس روند اسناد علمی فارسی انجام نشده است، ما *در این پژوهش قصد* ارائه روشی برای تحلیل‌روند و پیش‌بینی فناوری بر اساس اسناد علمی موجود در پایگاه گنج با استفاده از روش‌های یادگیری ماشین را داریم. استفاده از روش‌های آماری و داده‌های تاریخی باعث افزایش دقت عملکرد نسبت به روش‌های دیگر می‌شود. با ترکیب روش‌های یادگیری ماشین با نظرات متخصصان کمبودهای این روش‌ها کاسته می‌شود. از آنجا که تاکنون در حوزه تحلیل اسناد علمی در زبان فارسی پژوهش‌های زیادی انجام نشده است در استفاده از روش‌های یادگیری ماشین در این زمینه با چالش‌های زیادی روبرو خواهیم شد (ازجمله این چالش‌ها می‌توان به استخراج موضوع‌ها، خوشه‌بندی، دسته‌بندی، استخراج کلیدواژه‌ها اشاره کرد).

یافته‌های علمی گام نخست در شروع یک فناوری است (Anna Sokolova ۲۰۰۳Martino , ۲۰۰۵, Daim, Rueda, and Martin ۲۰۱۴)، برای ورودی سیستم ما از پارسا[۱] ثبت شده در سامانه گنج پژوهشگاه علوم و فناوری اطلاعات ایران استفاده خواهیم کرد. خروجی این سیستم روند تولید علمی که در حال حاضر در کشور در حال وقوع است و نمایی از آینده فناوری کشور را نشان می‌دهد می‌تواند در برنامه‌ریزی‌های آینده، تولید نقشه‌های علمی، چشم‌اندازهای وزارتخانه‌های و کنترل و پایش روند حرکت سازمان‌های مربوطه مورداستفاده قرارگیرد.

۱-۳- ضرورت و هدف پژوهش

سازمان‌های بزرگ اهداف و چشم‌اندازهای کوتاه‌مدت و بلندمدتی را برای خود متصور هستند و برای رسیدن به آن برنامه‌ریزی‌هایی انجام می‌دهند. برای مثال چشم‌اندازهای علم و فناوری و نقشه جامع علمی کشور ازجمله نمونه‌های این نوع برنامه‌ریزی‌ها است. تحلیل وضعیت فعلی و پیش‌بینی، ابزارهایی قدرتمند در فرایند برنامه‌ریزی هر کاری هستند. بنابراین درصورتی‌که از روش‌های تحلیل و پیش‌بینی استفاده نکنیم هیچ چشم‌اندازی نسبت به روند حرکت علمی کشور نخواهیم داشت و بدون توجه به نیازهای جامعه مسیری را برای حرکت انتخاب می‌کنیم که احتمال اینکه مسیر نادرستی باشد بسیار زیاد است. از طرفی روش‌های معمول پیش‌بینی که مبتنی بر نظر متخصصان است، دلایل متخصصان قابل استناد و مکتوب شدن نیست و این نظرات می‌توانند بر اساس تمایلات شخصی، پیش‌فرض‌های نادرست و یا حتی منافع شخصی اتخاذ شوند. با فرض اینکه پیش‌بینی انجام‌شده نیز درست باشد، می‌بایست روند حرکتی و وضعیت جاری جامعه علمی و فناوری کشور را مورد تحلیل و بررسی قرارداد و توسط روش‌هایی از صحت حرکت آن اطمینان حاصل کرد.

این در حالی است وجود یک سیستم که وظیفه تحلیل وضعیت جاری را دارد، با دیدی که نسبت به آینده به برنامه‌ریزان می‌دهد آن‌ها را در اتخاذ سیاست‌های درست یاری می‌کند. از طرفی باوجوداین سیستم دلایل سیاست‌های اتخاذشده قابل ثبت و نمایش هستند و نظرات و تمایلات شخصی در ارائه سیاست‌ها کمتر دخیل می‌شود و دقت بالاتر خواهد رفت. با استفاده از این سیستم

[۱] پایان‌نامه‌ها و رساله‌ها

هزینه‌های بسیار زیادی که مربوط به استفاده از متخصصان آن حوزه است کاهش پیدا خواهد کرد و سرعت اتخاذ تصمیمات نیز به میزان قابل‌توجهی کاهش پیدا می‌کند. مزیت دیگر استفاده از این سیستم این است که می‌توان روند حرکت جاری علمی و فناوری را با سیاست‌های اتخاذشده مقایسه و مورد تحلیل قرارداد. و حتی وضعیت جاری را می‌توان با وضعیت علمی دنیا مورد مقایسه قرارداد. با یک چنین سیستمی می‌توان موضوعاتی که به‌صورت مشترک در مراکز علمی در حال تحقیق است را تشخیص داده و سیاستی مناسبی را دراین‌باره اتخاذ کرد.

بنابراین **هدف اصلی** این پژوهش ارائه قالب کاری و سامانه‌ای برای تحلیل‌روند و رشد علم و فناوری بر اساس پارسا[1]ی ثبت شده دانشگاه‌ها در سامانه گنج است. ازآنجاکه رسیدن به هدف اصلی به‌یک‌باره ممکن نیست، آن را به اهداف میانی تقسیم می‌کنیم. در این رساله اهداف جانبی دیگری همچون موارد زیر نیز دنبال می‌کند:

۱. بررسی ساختار چیکده‌های پارساها؛
۲. راه‌حلی برای استخراج و پالایش کلمات‌کلیدی؛
۳. انتخاب روش مناسبی برای دسته‌بندی رساله‌ها؛
۴. انتخاب روش مناسبی برای خوشه‌بندی رساله‌ها؛
۵. پیش‌بینی روند تحقیقاتی بر اساس رساله‌های انجام‌شده.

بر اساس (Martino ۲۰۰۳, Anna Sokolova ۲۰۱۴) ازآنجاکه پروپوزال‌ها و یافته‌های علمی ازجمله علائم بروز یک فناوری هستند در این پژوهش ما تحلیل‌ها و پیش‌بینی‌ها را صرفاً بر اساس سامانه گنج (که شامل رساله‌ها و پروپوزال‌ها هستند). انجام خواهیم داد.

۱-۴- سؤالات پژوهش

[1] پایان‌نامه‌ها و رساله‌ها

پرسش اصلی که در این پژوهش مطرح است این است که آیا می‌توان روشی ارائه کرد که باعث شود روند علم و فناوری را تحلیل و پیش‌بینی کنیم؟ در صورتی که از استاندارد داده‌کاوی CRISP-DM برای پاسخ به این پرسش استفاده کنیم، به ترتیب پرسش‌های فرعی زیر را می‌بایست پاسخ داد:

۱. ساختار چیکده‌های پارساها به چه صورت است؟
۲. چه روشی برای استخراج و پالایش کلمات‌کلیدی مناسب است؟
۳. چه روشی برای دسته‌بندی رساله‌ها مناسبی است؟
۴. چه روشی برای خوشه‌بندی رساله‌ها مناسب است؟
۵. پیش‌بینی روند تحقیقاتی بر اساس رساله‌های انجام‌شده به چه صورت خواهد بود؟

۱-۵- روش انجام پژوهش

همانطور که اشاره شد برای انجام پژوهش از استاندارد CRISP-DM که یک ساختاری برای برنامه‌ریزی و اجرای پروژه‌های داده‌کاوی فراهم می‌کند استفاده شده است. این روش توسط بنیان فناوری اطلاعات و برنامه‌های راهبردی اروپا ارائه‌شده است. توضیحات بیشتر در خصوص انجام جزئیات این استاندارد در بخش ۲-۱-۲-۲ آورده شده است.

درگام نخست بعد از درک اهداف مسئله (پژوهش جاری) به مطالعات در خصوص روش‌های تحلیل‌روند علم و فناوری خواهیم پرداخت و نیازمندی‌های آن‌ها را استخراج می‌کنیم. در گام بعدی به شناخت و فهم داده‌هایی که قرار است مورد تحلیل و بررسی قراردهیم خواهیم پرداخت. برای این منظور می‌بایست داده‌های مورد نیاز استخراج و ساختار چکیده‌ها و کلیدواژه‌های پارسا را مورد بررسی و ارزیابی قراردهیم. بعد از آنکه متون پارسا مورد بررسی و تحلیل قرار گرفت در گام بعدی بر اساس مجموعه اسنادی که استخراج شده‌اند، مجموعه‌داده‌هایی تشکیل خواهند شد تا بر اساس این مجموعه‌داده‌ها بتوانیم عملیات داده‌کاوی و کشف الگور را انجام دهیم. در این گام در رساله روشی برای تشکیل مجموعه‌داده ارائه داده‌ایم. در گام چهارم با استفاده از مجموعه‌داده‌های تشکیل شده روشی برای استخراج و تشخیص موضوع‌ها علمی ارائه می‌دهیم تا تحلیل‌روند بر مبنای آن انجام شود. در گام پنجم نیز بخاطر ضعف در شاخص‌ها و روش‌های ارزیابی گذشته شاخص و روش‌هایی برای

ارزیابی موضوع‌ها و تحلیل‌روند در این رساله ارائه‌شده است. ساختار کلی روش پیش‌نهادی برای تحلیل‌روند به شکل زیر است.

شکل ۱-۱ مراحل تحلیل‌روند پیشنهادشده در این رساله

۱-۶- دانش‌افزایی[۱] (سهم‌های علمی)

[۱] Contributions

در این رساله برای نخستین بار روشی برای تحلیل‌روند با استفاده از یادگیری ماشین و به صورت تمام خودکار برای متون علمی فارسی ارائه‌شده است. علاوه بر آن در گام‌های تحلیل‌روند نوآوری‌هایی ارائه داده‌شده تا هر گام از تحلیل‌روند نسبت به روش‌های متداول بهینه عمل کند. **Error! Reference source not found.** مراحل مختلف تحلیل‌روند استفاده شده در این رساله را نشان می‌دهد. در صورتی که مراحل طرح ارائه‌شده را با استاندارد داده‌کاوی CRISP مطابقت دهیم شکل ۱-۲ سهم علمی رساله در گام‌های مختلف نشان داده‌شده است.

بعد از شناخت و بیان مسئله، نیازمندیم تا در گام بعدی به بررسی انواع داده‌های موجود بپردازیم. نوع و خصوصیات داده اهمیت فراوانی دارد و می‌تواند روش تحلیل را تحت تاثیر خود قرار دهد. از آنجا که داده‌های در دسترس فرداده‌های پارسای سامانه گنج است رابطه بین کلیدواژه‌های بکار گرفته‌شده توسط نمایه‌ساز و نویسنده، میزان مشارکت چکیده و عنوان در انتخاب کلیدواژه‌ها و منطبق با استاندارد بودن آن مورد بررسی و ارزیابی قرار گرفت. از اطلاعات استخراج شده از این بخش می‌توان در نحوه انتخاب و استخراج کلیدواژه‌ها از فرادادهای پارسا بهره گرفت.

بعد از شناخت داده‌ها می‌بایست متون برای پردازش و تحلیل آماده شوند. نحوه انتخاب کلیدواژه‌ها، استخراج ویژگی و نحوه بازنمایی کلمات (که مبنای استخراج موضوع‌ها و تحلیل‌روند هستند) از جمله ضروریات این بخش هستند که برای برای هر کدام از آن‌ها روشی جدید در این رساله ارائه گردیده است.

بنابراین در گام آماده‌سازی داده‌ها بعد از اجرای متداول روش‌های پاک‌سازی داده‌ها، رابطه‌ای برای انتخاب تعداد کلیدواژه‌های موثرتر از کلیدواژه‌های کاندید ارائه‌شده است. سپس با استفاده از کلیدواژه‌های انتخاب‌شده، شاخصی برحسب احتمال هم‌رخدادی دو کلیدواژه در یک چکیده با هدف تحلیل‌روند ارائه‌شده است. با استفاده از این شاخص مجموعه‌داده‌هایی تشکیل‌شده است تا با استفاده از آن بتوان موضوع‌های موجود در حوزه‌های علمی را استخراج کرده و روند آن‌ها را تحلیل کنیم.

هدف از گام چهارم که مدل‌سازی نام دارد تشخیص موضوع‌ها و محاسبه شاخص‌های تحلیل‌روند است. تحلیل‌روند با استفاده از روابط استخراج شده بین موضوعی از طریق قواعد انجمنی، رسم نمودار

راهبردی برحسب ماتریس هم‌رخدادی موضوعی، محاسبه شاخص‌های مرکزیت و بلوغ بر اساس روابط موضوع‌ها از دید اسناد پژوهشی، ارائه روشی برای خوشه‌بندی با خلوص بالا برای تفکیک غیرنظارت شده اسناد، بررسی و ارزیابی روش‌های دسته‌بندی اسناد علمی فارسی از دانش‌افزایی‌های مرتبط با این بخش هستند.

شکل ۱-۲ سهم علمی رساله در گام‌های مختلف

فصل ۲- مبانی نظری و پیشینه‌پژوهش

۲-۱- مبانی نظری

در این بخش به بررسی مبانی نظری و مفاهیم مورد استفاده در پژوهش خواهیم پرداخت. این مفاهیم در گام‌های پژوهش و پیاده‌سازی آن مورد استفاده قرار گرفته‌اند.

۲-۱-۱- مقدمه

رشد روزافزون فناوری‌ها در صنایع تحولات زیادی ایجاد می‌کند و ازاین‌رو فرصت‌ها و چالش‌های جدیدی در جامعه و کسب‌وکار پدید می‌آید. همه مدیران از روند رشد این تحولات آگاه‌اند؛ اما پرسش مهمی که مطرح می‌شود این است که فناوری به کدام سمت‌وسو در حال حرکت است؟ روند آن چگونه است؟ و تحولات کوچک و بزرگ چه هنگامی رخ می‌دهند؟

دانش در رابطه با وضعیت فعلی و پیش‌بینی آینده برای حفظ بقا لازم و ضروری است. برای نمونه همه ما در ابتدای روز با بررسی وضع هوا لباس مناسب را به تن می‌کنیم و یا شرکت‌ها و سازمان‌های بزرگ و کوچک امروزی که با تحول بزرگ فناوری روبرو هستند نیز می‌بایست برای حفظ بقا، افقی از آینده فناوری را برای خود متصور شوند. این پیش‌بینی در همه سطوح تصمیم‌گیری قابل‌استفاده است، چه در سطح تصمیمات ملی و چه در سطح بنگاه‌های کوچک اقتصادی (Firat, Woon, and Madnick ۲۰۰۸). ازجمله کاربردهای پیش‌بینی فناوری می‌توان به اولویت‌بندی در بخش تحقیق و توسعه، برنامه‌ریزی در خصوص توسعه محصولات جدید، وضع تصمیم‌های راهبردی برای صدور و یا دریافت مجوزهای فناوری اشاره کرد (Jun, Park, and Jang ۲۰۱۲).

از طرفی فناوری‌های نوظهور باعث توسعه نوآوری‌هایی در زمینه‌های گوناگون هستند (Kim, Suh, and Park ۲۰۰۸). توسعه یک فناوری، وابسته به ظهور فناوری‌های مربوط به آن هست

(Bengisu and Nekhili ۲۰۰۶, Daim et al. ۲۰۰۶). ازاین‌رو استخراج رابطه بین یک فناوری هدف با دیگر فناوری‌های وابسته به آن نیز برای توسعه فناوری بسیار حائز اهمیت است. استخراج این رابطه را پیش‌بینی فناوری نوظهور گویند که محدوده‌های ظهور فناوری را نشان می‌دهد (Jun and Lee۲۰۱۲). پیش‌بینی فناوری اشکال گوناگون و متفاوتی را به خود می‌گیرد و در حوزه‌های مختلفی کاربرد دارد که می‌توان به پیش‌بینی مطالعات ملی و نقشه جامع راه فناوری، هوش فناوری، هوش رقابتی و هوش فناوری رقابتی اشاره کرد (.Coates et al۲۰۰۱).

برای پیش‌بینی روش‌هایی چون نظرات متخصصان، پایش و نظارت بر اطلاعات، روش‌های آماری، برون‌یابی روند، مدل‌سازی، سناریونویسی ارائه‌شده است (Firat, Woon, and Madnick ۲۰۰۸). برخی از این روش‌ها بر روی داده‌های گذشته عمل می‌کنند و برخی دیگر نیز بر اساس نیازهایی که در آینده احساس می‌شوند، طراحی شده‌اند. روش‌های ترکیبی همچون درخت ارتباط (Esch۱۹۶۵) و سناریونویسی روش‌هایی هستند که بر روی هر دو نوع داده تحلیل می‌کنند (Gordon and Helmer۱۹۶۴). در تقسیم‌بندی دیگر می‌توان روش‌ها را ماشینی، انسانی و ترکیبی دانست.

اولین بار پیش‌بینی فناوری به‌صورت نظام‌مند در سال ۱۹۳۵ توسط کمیسیون منابع معاملاتی جدید ایالات‌متحده به اجرا درآمد (Ogburn۱۹۳۷) که انتشار گزارشات نشان می‌دهد که نگرانی آن زمان بیشتر بر روی تغییرات فناوری بوده است (Schnaars, Chia, and Maloles III۱۹۹۳). بعد از جنگ جهانی مدلی ارائه شد که در آن نوآوری‌هایی که جریان فناوری[۱] را از علوم پایه تا تولید محصول و تجاری‌سازی نقش دارند را نشان می‌داد (Bush۱۹۴۵). بعد از جریان «اسپانتیک[۲]» ایالات‌متحده اولین استراتژی مدیریت تحقیق و توسعه در صنعت وضع کرد و بعد از آن پیش‌بینی فناوری به‌عنوان ابزاری نظام‌مند برای کشف فناوری‌های آینده برای توسعه جنگ‌افزارهای نسل‌های آینده به اجرا درآمد (Roussel۱۹۹۱).

[۱] technology flow

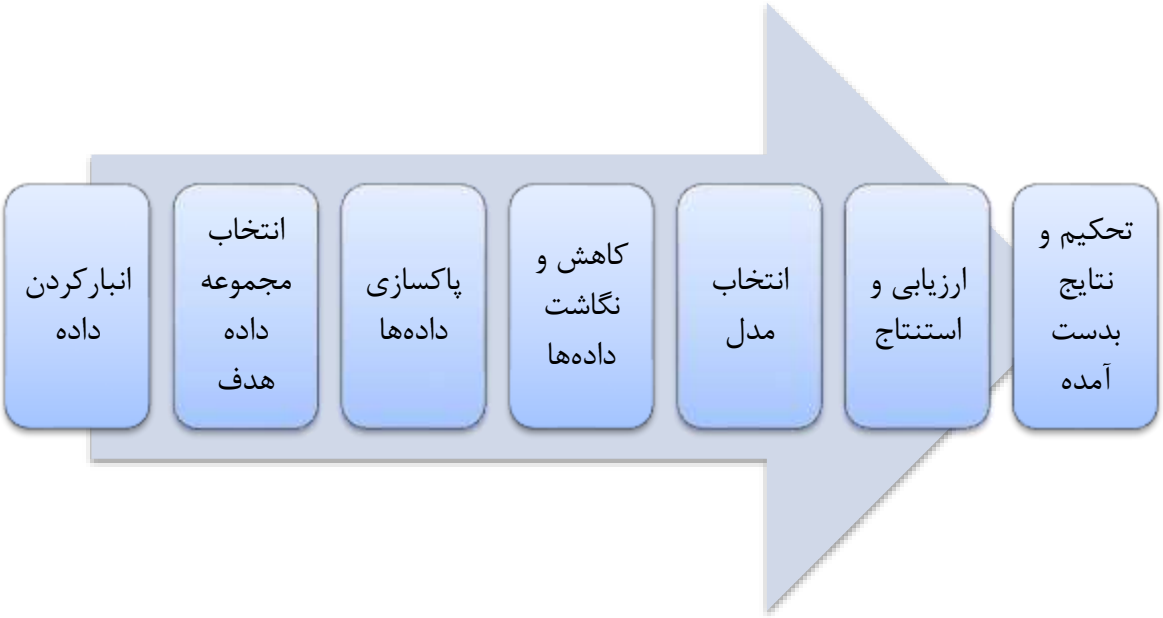
[۲] SPANTIK

مدیران و تصمیم‌گیرندگان سازمان‌ها اغلب بر اساس حوزه دانشی که در اختیار دارند تصمیم‌های خود را اتخاذ می‌کنند؛ بنابراین دانش محدود سیاست‌گذار، پیش قضاوت‌ها، هزینه‌بر بودن استفاده از متخصص‌ها و عدم ثبت دلایل برای اتخاذ تصمیم، نیاز به یک فرایند برای پیش‌بینی فناوری آینده که ازنظر دقت، سرعت و هزینه این کاستی‌ها را جبران کند، احساس می‌شود. روش‌های کتابشناختی ازجمله روش‌های ترکیبی هستند که با استفاده از سامانه‌های رایانشی به تجزیه‌وتحلیل منابع و مدارک موجود پرداخته و با استفاده از روش‌های گوناگونی ازجمله یادگیری ماشین و یا آماری به پیش‌بینی آینده می‌پردازند. این روش‌ها می‌توانند به‌طور کاملاً خودکار به‌صورت ماشینی و یا ترکیب ماشین و متخصصان آن حوزه انجام شود. روش‌ها و قالب‌های کاری متعددی در این زمینه ارائه‌شده که در بخش پیشینه پژوهش به آن پرداخته خواهد شد.

در پیش‌بینی فناوری، «فناوری» ابزارها، شیوه‌ها و فرایندهایی است که برای برآورده کردن نیازهای انسان مورداستفاده قرار می‌گیرد (Martino۱۹۹۳). پیش‌بینی فناوری را به‌طورکلی هرگونه روش هدفمند و نظام‌مند برای پیش‌بینی و فهم در زمینه‌های نرخ رشد فناوری، جهت رشد فناوری، ویژگی‌های رشد آن و تأثیرات تغییرات فناوری (بخصوص ابداعات، نوآوری‌ها) خوانده‌اند (Firat, Woon, and Madnick۲۰۰۸). پیش‌بینی فناوری عبارت است از پیش‌بینی تمایلات و روند فناوری مانند جهت رشد و نرخ توسعه فناوری (Jun and Lee۲۰۱۲, Jun ۲۰۱۱, Porter ۱۹۹۱a).

۲-۱-۲- کشف دانش و داده‌کاوی

منظور استخراج دانش از پایگاه داده‌ها، کشف اطلاعات مفید از پایگاه داده است (Fayyd, Shapiro, and Smyth۱۹۹۶). داده‌کاوی یک مجموعه از شیوه‌هایی است که به‌طور خودکار در مجموعه پایگاه داده‌های بزرگ جستجو و روابط پیچیده را پیداکرده و نشان می‌دهد. داده‌کاوی فرایند تحلیل حجم زیادی از داده‌های و استخراج اطلاعات مفید از آن داده‌ها و به دست آوردن معنا از آن‌هاست؛ بنابراین داده‌کاوی داده‌ها را به اطلاعات مفید و قابل‌استفاده تبدیل می‌کند. هدف از داده‌کاوی را می‌توان کشف دانش پنهان‌شده از داده‌های فراوانی است که به از دید انسان پنهان هستند. گام‌هایی که در فرایند کشف دانش مورداستفاده قرار می‌گیرند به‌قرار زیر هستند:

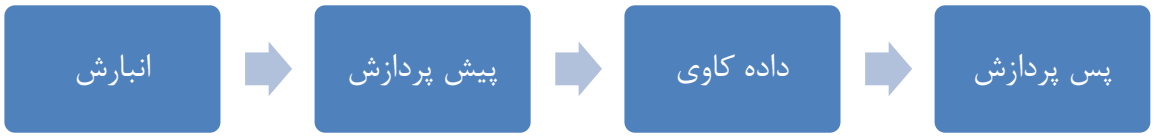


شکل ۲-۱ مراحل کشف دانش (Fayyd, Shapiro, and Smyth ۱۹۹۶).

- انبارش داده: در این بخش این امکان فراهم می‌شود که داده‌ها از محیط‌ها و منابع مختلف به اشکال گوناگون جمع‌آوری شوند.
- انتخاب مجموعه‌داده هدف: در این بخش یک مجموعه‌داده ساخته می‌شود.
- پاک‌سازی داده‌ها: داده‌های نویز و پرت را از مجموعه‌داده حذف می‌کند.
- کاهش و نگاشت داده‌ها: بر طبق نیاز و کاربرد، داده‌ها از بین مجموعه داده‌ای آماده‌شده انتخاب و به شکل قابل‌پردازشی تبدیل خواهند شد.
- انتخاب مدل: این بخش یک الگوریتم مناسب را برحسب کاربرد انتخاب خواهد کرد.
- داده‌کاوی: به دنبال یک الگوی مفید و موردنظر خواهد گشت.
- ارزیابی و استنتاج: نتایج به‌دست‌آمده شده در این بخش ارزیابی و اعتبار سنجی خواهد شد.
- تحکیم نتایج به‌دست‌آمده: اطلاعات کشف شده در این گام برای استفاده مجدد و بررسی آماده می‌شوند.

- هنجارسازی[1]: در عمل هنجارسازی نوشتار، متن دلخواه ورودی بدون تغییر در معنای آن به ساختار قانونی و مطابق با پردازشگر متن تبدیل می‌شود. برای مثال حروف و کلمات (مانند ک، گ، ی) را به شکل یکسان بازنویسی می گردد یا اینکه علائم نگارشی، حروف، فاصله بین کلمات و غیره به شکل یکسانی نوشته خواهند شد. لازم به ذکر است که بدون انجام هنجارسازی، ورودی‌هایی به‌ظاهر متفاوت اما در معنی یکی به پردازشگر داده‌شده که این باعث ایجاد خطا در تحلیل‌ها توسط پردازش‌های انجام‌شده خواهد شد.

انتخاب ویژگی تمام مراحل کشف دانش را می‌تواند تحت تأثیر خود قرار دهد. البته در بعضی از الگوها مختلط برخی از این مراحل با یکدیگر ترکیب شده و اشکال متفاوتی نشان داده می‌شوند. شکل ۲-۲ نمونه‌ای از این الگوها است. در این الگو گام‌های کشف دانش را به‌صورت خلاصه به چهار گام اصلی تقسیم کرده است که عبارت‌اند از: انبارش داده، پیش‌پردازش (۲ و ۳ و ۴)، داده‌کاوی (۵ و ۶) و پس پردازش (۷ و ۸) تقسیم کرد.



شکل ۲-۲ مدل فشرده شده مراحل کشف دانش (Fayyd, Shapiro, and Smyth ۱۹۹۶).

در گام پیش‌پردازش بر روی داده‌ها به‌منظور آماده‌سازی برای مرحله بعد پردازش‌هایی انجام خواهد شد. ازآنجاکه ویژگی‌های زیاد برای داده‌ها، باعث مشکل شدن در پردازش و تحلیل آن‌ها می‌شود و از طرفی دیگر خواندن آن‌ها به یک‌باره برای رایانه مشکل است، کاهش ابعاد امری ضروری است (Liang et al. ۲۰۱۲). این نکته نیز باید یادآوری شود که داده‌ها در مرحله انبار داده فقط

[1] Text Normalization

به‌منظور داده‌کاوی برای یک هدف خاص جمع‌آوری نشده‌اند و نیاز است که ویژگی‌ها و داده‌های مرتبط با هدف را انتخاب کنیم.

در فاز دوم از الگوریتم‌های گوناگونی می‌توان استفاده کرد. مسائل مانند انتخاب الگوریتم و مدل یادگیری، سرعت در رسیدن به جواب (سرعت در یادگیری) و نمایش مقدماتی نتایج در این بخش قابل بررسی هستند. در انتها عملیات پس پردازش انجام خواهد شد. ازآنجاکه ممکن است در نتایج، اطلاعات بسیار زیاد و یا موارد بدیهی وجود داشته باشد، در این فاز مرتب‌سازی و سازمان‌دهی داده‌های بخش داده‌کاوی صورت می‌پذیرد. ارزیابی نتایج به‌دست‌آمده ازجمله موارد موردبررسی در این فاز است. ازآنجا که قوانین مستخرج شده با پارامترها (ویژگی‌های داده‌ها) بیان می‌شوند، می‌توان نتیجه گرفت که هرچه ویژگی‌های داده‌ها کمتر باشند مدل آموزش دیده شده ساده‌تر و قابلیت درک پذیری بیشتری خواهد داشت.

متدولوژی‌های شناخته‌شده‌ای که برای داده‌کاوی مورداستفاده قرار می‌گیرند CRISP-DM[۱] و SEMMA[۲] نام دارند که در ادامه به بررسی آن‌ها خواهیم پرداخت.

SEMMA -۱-۲-۱-۲

متدولوژی SEMMA مخفف کلمات نمونه، کشف، تغییر، مدل و ارزیابی است. این متدولوژی توسط موسسه SAS توسعه داده‌شده است که یکی از بزرگ‌ترین تولیدکنندگان نرم‌افزارهای هوش تجاری آماری است. متدولوژی مذکور برای راهنمایی در خصوص پیاده‌سازی برنامه‌های داده‌کاوی مورداستفاده قرار می‌گیرد.

هرچند که CRISP-DM توسط بنیان فناوری اطلاعات و برنامه‌های راهبردی اروپا به‌منظور ارائه یک روش خنثی برای پروژه‌های داده‌کاوی ارائه‌شده است، شرکت SAS نیز متدولوژی خود را ارائه کرده است. این متدولوژی در کنار CRISP-DM در رده بهترین روش‌ها برای داده‌کاوی مدنظر قرارگرفته‌اند. فازهای این روش به شرح زیر هستند:

---

[۱] Cross-industry standard process for data mining

[۲] Sample, Explore, Modify, Model, and Assess

- **نمونه:** فرایند با نمونه‌گیری از داده‌ها شروع می‌شود.
- **اکتشاف:** تحلیل اکتشافی بر روی متغیرها به‌منظور پیدا کردن روابط احتمالی بین آن‌ها و جستجو داده‌های غیر نرمال در بینشان در راستای رسیدن به اهداف. این فاز می‌تواند با کمک ابزارهای بصری سازی صورت بگیرد.
- **تغییر:** این فاز شامل روش‌هایی برای انتخاب، ساخت و انتقال متغیرها در راستای آماده‌سازی آن‌ها برای مدل‌سازی داده‌ها انجام می‌شود.
- **مدل:** تمرکز این فاز برای بکار بردن شیوه‌های مدل‌های متنوع (داده‌کاوی) بر روی متغیرهای آماده شده است. به‌عبارت‌دیگر در این فاز مدلی تشکیل خواهد شد که خروجی مطلوب را بدهد.
- **ارزیابی‌ها:** آخرین فاز ارزیابی است. ارزیابی نتایج مدل برای نشان دادن قابلیت اطمینان و مفید بودن مدل ساخته‌شده انجام می‌گیرد.

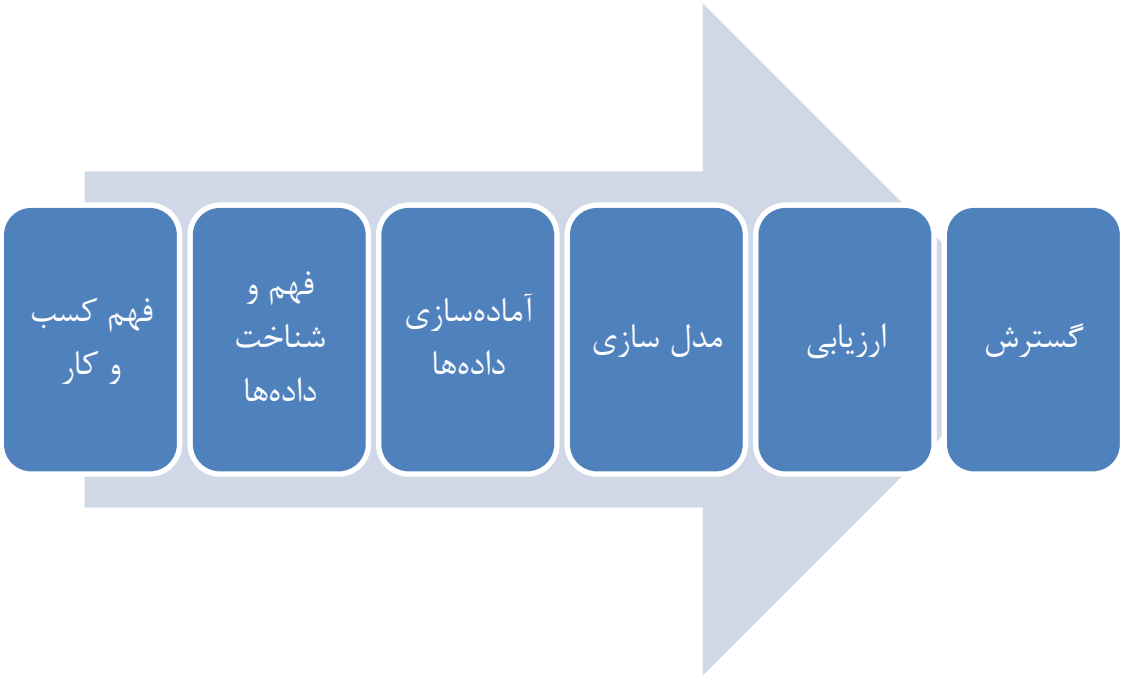
مشکلی که روش SEMMA دارد عدم توجه کافی از دیدگاه کسب‌وکار است و تمرکز این روش بیشتر بر روی داده‌کاوی است.

۲-۱-۲-۲- روش پژوهش CRISP-DM

روش پژوهش CRISP-DM یک ساختاری برای برنامه‌ریزی و اجرای پروژه‌های داده‌کاوی فراهم می‌کند. این روش توسط بنیان فناوری اطلاعات و برنامه‌های راهبردی اروپا ارائه‌شده است. روش CRISP-DM در اصل یکی از روش‌های تحلیلی متفاوت برای فرایند داده‌کاوی است که بسیار انعطاف‌پذیر است. مراحل این روش به‌قرار زیر است(Shearer ۲۰۰۰):

۱- فهم کسب و کار
۲- فهم و شناخت داده‌ها
۳- آماده‌سازی داده‌ها
۴- مدل‌سازی
۵- ارزیابی

۶- گسترش



شکل ۲-۳ مراحل مدل CRISP-DM

۲-۱-۲-۲-۱- فهم کسب و کار

مرحله اول از فرایند CRISP-DM فهم این است که از دیدگاه کسب‌وکار شما بدانید که چه‌کاری را می‌خواهید به اتمام برسانید. سازمان می‌تواند اهداف رقابتی و محدودیت‌هایی داشته باشد که باید به‌صورت مناسبی در تعادل باشند. هدف از این بخش، کشف فاکتورهای مهمی است که می‌توانند در خروجی اثر بگذارند. حذف این هدف می‌تواند در انتها باعث به دست آمدن خروجی اشتباهی بشود. در این بخش باید مشخص شود که خروجی مطلوب چه چیزی است؛ بنابراین باید اهداف مشخص شود، برنامه تولید پروژه تشکیل شده و معیارهای موفقیت کسب‌وکار مشخص می‌شوند.

۲-۱-۲-۲-۲- فهم و شناخت داده

در مرحله دوم از روش CRISP-DM شما باید داده‌های مورد نیاز مرحله قبل را فراهم کنید. در این مرحله شما می‌بایست با بررسی داده‌های خود مطمئن شوید که نیاز شما را برآورده می‌کنند.

درصورتی‌که از چند منبع داده‌ای استفاده می‌کنید باید این اطمینان را حاصل کنید که در زمان مناسبی آن‌ها را به‌صورت مناسبی ادغام کنید. گام‌هایی را که در این مدل پیشنهاد می‌شود به‌قرار زیر است:

- توصیف کردن داده
- اکتشاف داده
- تأیید کیفیت داده
- گزارش کیفیت داده

۲-۱-۲-۲-۳- آماده‌سازی داده

در مرحله قبل داده‌های خام دریافت شد و بعد از تأیید کیفیت به این مرحله رسیده است. در این مرحله دادهای اولیه‌ای که از مرحله قبل دریافت شد را با اعمال پالایش‌هایی انتخاب خواهیم کرد. در ادامه با استفاده از داده‌های آماده‌شده، یک مجموعه‌داده تشکیل می‌شود تا با استفاده از آن مدل‌های تحلیلی ساخته شوند. گام‌هایی را که در این مدل پیشنهاد می‌شود به‌قرار زیر است:

- انتخاب داده‌ها
- پاک سازی داده‌ها
- ساخت داده‌های مورد نیاز
- یکپارچه‌سازی داده‌ها

۲-۱-۲-۲-٤- مدل‌سازی

در این مرحله تحلیلگران روش‌هایی متناسب با نیازهای خود به‌منظور انجام تحلیل‌های آتی انتخاب می‌کنند. این روش‌ها بر روی داده‌هایی که از مرحله قبل به این مرحله آورده شده است اعمال می‌شوند. هدف این فاز ساختن مدلی است که قادر به توصیف مجموعه‌داده‌ها باشد. با استفاده از مدل ساخته‌شده ارزیابی و پیش‌بینی‌های مربوط به کسب‌وکار انجام خواهد شد. گام‌هایی را که در این مدل پیشنهاد می‌شود به‌قرار زیر است:

- انتخاب روش مدل‌سازی
- تولید طراحی آزمون

- ساخت مدل
- ارزیابی مدل

۲-۱-۲-۲-۵- ارزیابی

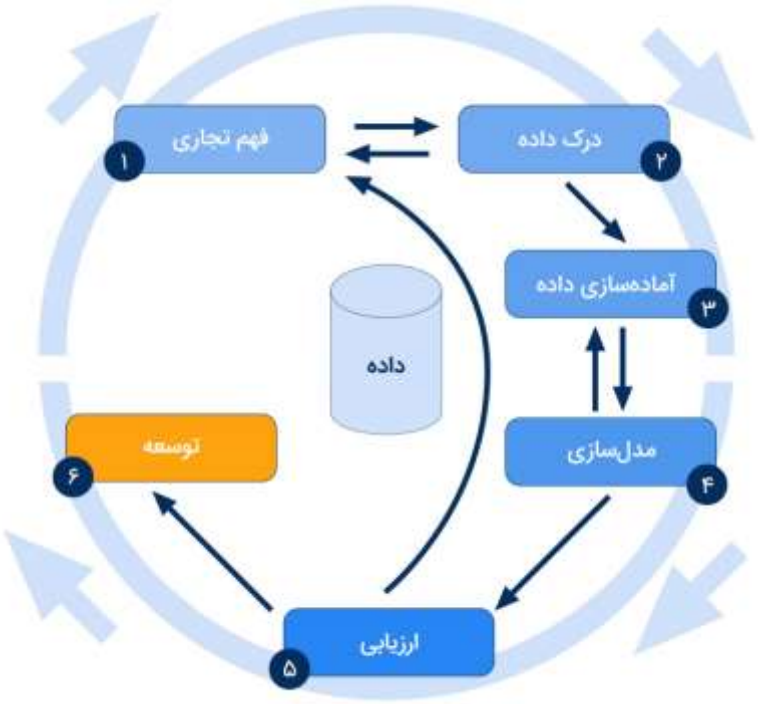
پس از به دست آمدن مدل، می‌بایست از صحت و درستی آن اطمینان حاصل شود؛ بنابراین می‌بایست مدل تشکیل‌شده تفسیر بشود. درصورتی‌که نتایج حاصل از ارزیابی رضایت‌بخش نبود، فازهای قبل تکرار می‌شوند تا نتایج مطلوبی به دست آید.

- ارزیابی نتایج
- بازنگری فرایند
- تعیین و مشخص کردن گام بعدی

۲-۱-۲-۲-٦- بکارگیری

بعد از تشکیل مدل و اطمینان از صحت عملکرد آن، می‌توان مدل را در فرآیندهای تجاری سازمان مورداستفاده قرار داد. از طرفی دیگر مدل تشکیل‌شده می‌تواند در مدت‌زمان مورداستفاده با داده‌های جدید به‌روز شود تا جامعیت خود را حفظ کند.

- برنامه ریزی توسعه
- نگهداری و پایش برنامه
- تولید گزارش نهایی
- مطالعه مجدد پروژه



شکل ۲-٤ روش پژوهش و استانداردCRISP-DM (Shearer ۲۰۰۰)

۲-۱-۲-۳- برخی از ابزارهای مدل‌سازی

۲-۱-۲-۳-۱- دسته‌بندها

الگوریتم‌های داده‌کاوی اهداف گوناگونی را می‌توانند دنبال کنند ازجمله این کاربردها می‌توان به دسته‌بندی، خوشه‌بندی و پیش‌بینی اشاره کرد. در دسته‌بندی، داده موجود به یکی از دسته‌هایی که از قبل مشخص شده‌اند تخصیص داده می‌شود. برای رسیدن به این هدف، ابتدا می‌بایست، یک دسته‌بند را آموزش داد؛ بنابراین مجموعه‌داده‌ها را به دو دسته داده‌های آموزش (برای یادگیری) و داده‌های تست (برای ارزیابی) تقسیم خواهیم کرد. برای آموزش دسته‌بندی هر داده‌ی مجموعه‌داده دارای برچسبی است که مشخص می‌کند به کدام دسته تعلق دارد. به این‌گونه الگوریتم‌هایی که از داده‌های برچسب خورده استفاده می‌کنند، الگوریتم‌های نظارت‌شده می‌گویند. درصورتی‌که داده‌های آموزش برچسب گزاری نشده باشند، الگوریتم‌های غیر نظارت (برای خوشه‌بندی) شده مورداستفاده واقع می‌شوند (Fayyad et al. ۱۹۹۶).

الگوریتم‌های دسته‌بندی را می‌توان بر اساس ویژگی‌ها به دو صورت تنبل[1] و مشتاق[2] دسته‌بندی کرد (Aha۱۹۹۸). در دسته الگوریتم‌های تنبل انتخاب نتیجه خروجی به داده‌های قبلی بستگی دارد و می‌بایست برای محاسبه نتیجه آن‌ها نیز در دسترس باشند بدین دلیل به آن‌ها الگوریتم‌های یادگیری موردی، مبتنی بر نمونه یا مبتنی بر حافظه نیز گفته می‌شود. نمونه این الگوریتم KNN است (Cover and Hart۱۹۶۷). آموزش‌های تنبل از آن جهت به این صورت نام گزاری شده‌اند که این الگوریتم ها راه حلی کلی برای داده‌ها آموزش نمی‌دهد و تا زمانی که نیاز به یک کلاس برای داده جدید نباشد از داده‌های آموزش استفاده نمی‌کند. در مقابل الگوریتم‌های مشتاق از داده‌های آموزش یک مفهوم توصیفی ارائه می‌دهند که به‌طورکلی در مقابل داده‌های آموزش بسیار فشرده‌تر هستند.

به‌طورخلاصه فرایند دسته‌بندی شامل سه مرحله است: گام اول، آموزش کلاسبند، گام دوم، کلاسبندی داده‌ها و گام سوم اندازه‌گیری کارایی کلاسبند. بعد از آموزش، کلاسبندها قابلیت تقریب تابع اصلی حالت یک رخداد را *به دست می‌آورند*. دو پارامتر داده و الگوریتم یادگیری در کارایی تأثیر زیادی دارند. برای مثال اگر داده‌ها زیاد باشند، ممکن است کامپیوتر قابلیت هندل کردن آن‌ها را نداشته باشد و از طرفی اگر داده‌ها اندک باشند، ممکن است الگوریتم یادگیری قابلیت یادگیری مطلب با معنی را نداشته باشد. داده‌های با نویز و دارای خطا خود باعث اشتباه در الگوریتم یادگیری خواهند شد. ازآنجاکه الگوریتم‌های طبقه‌بندی در حالات مختلف دارای کارایی و دقت متفاوتی هستند، در مسئله انتخاب ویژگی ما باید محدودیت‌های طبقه بند و داده‌ها را در نظر بگیریم.

۲-۱-۲-۳-۱-۱- انتخاب ویژگی برای دسته‌بندی

برای انتخاب ویژگی می‌بایست به سؤال‌هایی همچون: چگونه جستجو برای بهترین ویژگی‌ها انجام شود، چه معیاری برای انتخاب بهترین ویژگی در نظر گرفته شود، چگونه ویژگی‌ها انتخاب، حذف یا اضافه شوند، پاسخ داد. درصورتی‌که بنا به دلایلی تعداد ویژگی‌های ما زیاد باشند ممکن است اشکالات زیر به وجود آیند:

---

[1] Lazy

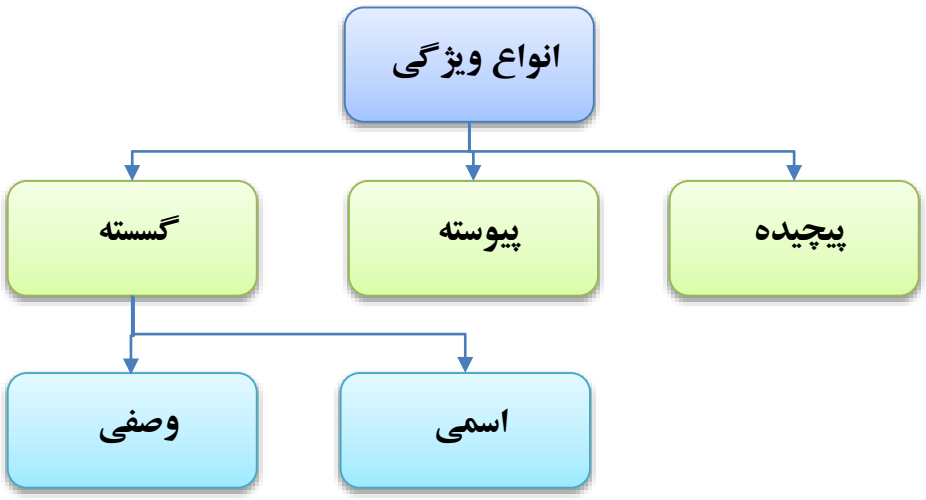
[2] Eager

- مجبور می‌شویم نمونه‌های مجموعه‌داده را برای در نظر گرفتن همه حالات بالا ببریم.
- ویژگی‌های نامربوط می‌توانند الگوریتم یادگیری را به مسیری نادرست هدایت کنند و دقت آن طبقه‌بند کاهش یابد.

با افزایش ویژگی‌ها طبقه‌بند مجبور است پیچیدگی یادگیری و روابط خود را افزایش دهد. راسل و نورویژ در کتاب خود گفته‌اند که یک طبقه‌بند ساده دقت بیشتری نسبت به نوع پیچیده خود دارد؛ بنابراین برای افزایش دقت در طبقه‌بندی می‌بایست از یک زیرمجموعه حداقلی از ویژگی‌ها استفاده کرد، به‌طوری‌که دقت طبقه‌بند کاهش پیدا نکند. از طرفی با کاهش ویژگی‌ها به مدلی با پیچیدگی کمتر دست پیدا خواهیم کرد که قابل‌فهم‌تر خواهد بود. این مدل ساده کلی‌تر بوده و نسبت به مدل خاصی که از همه ویژگی‌ها استفاده می‌کند مسئله بیش برازش را در خود ندارد؛ بنابراین مسئله انتخاب ویژگی برای کلاس‌بندی بسیار لازم و ضروری خواهد بود تا بتوانیم آن را به‌درستی آموزش دهیم و از آن استفاده کنیم.

۲-۱-۲-۳-۲- انواع روش‌های انتخاب ویژگی

هر ویژگی می‌تواند یک مقدار گسسته و یا پیوسته انتخاب کند. مقادیر پیوسته می‌توانند اعداد صحیح و نامحدودی باشند اما در مقابل مقادیر گسسته تعدادی متناهی هستند. مقادیر گسسته خودشان می‌توانند به دو دسته، ترتیبی و اسمی شکسته شوند. در ویژگی‌های ترتیبی، خاصیت مرتب بودن در یک جهت و یا در دو جهت وجود دارد، به این صورت که می‌توان ویژگی‌ها را در راستای آن مرتب کرد برای مثال افراد را می‌توان بر اساس سن مرتب کرد؛ اما در ویژگی‌های اسمی دیگر این خصوصیت وجود ندارد و هیچ‌گونه ترتیبی بین آن‌ها به‌طور طبیعی وجود ندارد. به‌عنوان نمونه نمی‌توان اشیاء را بر اساس رنگ‌های قرمز، آبی و سبز به‌صورت ذاتی مرتب کرد. به‌صورت کلی می‌توان ویژگی‌های پیوسته را به گسسته و بالعکس تبدیل کرد.



شکل ۲-۵سلسله‌مراتب انتخاب ویژگی (Fayyd, Shapiro, and Smyth ۱۹۹۶).

۲-۱-۲-۳-۲-۱- جستجو

در انتخاب ویژگی قرار این است که یک مجموعه‌ای از ویژگی‌هایی که برای کلاس‌بندی بهینه هستند را پیدا کنیم. فرض کنید که هیچ اطلاعاتی در مورد این‌که این مجموعه بهینه کجا قرار دارند نداریم. راه‌های متفاوتی برای پیدا کردن این مجموعه بهینه ارائه‌شده است که به‌اختصار در زیر ملاحظه خواهیم کرد:

- **تولید ویژگی به‌صورت مستقیم:** در این حالت از یک مجموعه خالی شروع کرده و هر بار یک ویژگی که ازنظر معیار بهترین ویژگی ما است را به مجموعه اضافه خواهیم کرد تا به حالت بهینه‌ای که مدنظر داریم برسیم (Ververidis and Kotropoulos ۲۰۰۸).
- **تولید ویژگی به‌صورت معکوس**: در این حالت از یک مجموعه کامل ویژگی‌ها شروع کرده و هر بار یک ویژگی را که ازنظر معیار بدترین ویژگی ما است را از مجموعه حذف خواهیم کرد تا به حالت بهینه‌ای که مدنظر داریم برسیم (Jain and Zongker ۱۹۹۷).
- **تولید دو جهته:** در این حالت از دو مجموعه مجزا استفاده خواهیم کرد که در یکی به‌صورت مستقیم و در مجموعه دیگر به‌صورت معکوس عمل خواهد شد.

- **تولید ویژگی به‌صورت تصادفی:** در این حالت در هر مرحله هم می‌توانیم ویژگی به مجموعه اضافه و هم می‌توانیم حذف کنیم؛ اما این حذف و اضافه به‌صورت تصادفی خواهد بود.

۲-۱-۲-۳-۲-۲- تک متغیره و چند متغیره

تمرکز در این دیدگاه بر روی ارتباط بین ویژگی‌ها است، برای مثال این‌که ما یک ویژگی یا چند ویژگی را در هر مرحله اضافه و یا کم کنیم (Guyon and Elisseeff ۲۰۰۳). برای روشن شدن این موضوع که برای انتخاب ویژگی‌ها از چه روشی استفاده می‌شود سه روش درخت تصمیم، نایو بیز[۱] و شبکه عصبی را در نظر بگیرید. در نایو بیز آمار و احتمالات پیشین و پسین کلاس‌های تک‌تک ویژگی‌ها استخراج می‌شود. در درخت تصمیم، خاصیت ویژگی‌ها در خصوص تقسیم‌پذیری یک گره درخت تصمیم موردبررسی قرار می‌گیرند، اما یک شبکه عصبی از تمام ویژگی‌ها در طول یک ایپوک برای آموزش شبکه خود استفاده خواهد کرد. به‌طور خلاصه، دو کلاس بند اول تک متغیره و کلاس بند دوم چند متغیره است (Liu and Motoda ۲۰۱۲, Chandrashekar and Sahin ۲۰۱۴).

۲-۱-۲-۳-۲-۳- مدل نظارت شده و غیر نظارت شده

همان‌طور که در بخش‌های قبلی ذکر شد، روش‌های نظارت‌شده به روش‌هایی گفته می‌شود که داده‌های آموزشی دارای برچسب باشند و درصورتی‌که داده‌های آموزشی برچسب نداشته باشند به آن روش غیر نظارت‌شده گفته می‌شود. در الگوریتم‌های انتخاب ویژگی نیز به‌صورت مشابه، اگر از ویژگی برچسب دار بودن مجموعه‌داده استفاده شود به آن روش نظارت‌شده و در غیر این صورت، غیر نظارت‌شده خوانده می‌شوند.

در هرکدام از این روش‌ها الگوریتم انتخاب ویژگی به دو دسته فیلترینگ[۲] و پوششی[۳] تقسیم می‌شوند (Kohavi and John ۱۹۹۷, Das ۲۰۰۱). در روش پوششی بعد از انتخاب یک زیرمجموعه

[۱] Naïve Bayes

[۲] Filtering

[۳] Wrapper

از ویژگی‌ها، ارزش آن ویژگی‌ها را با ارزیابی الگوریتم داده‌کاوی مورداستفاده می‌سنجیم، درصورتی‌که دقت بهتر شد ویژگی از ویژگی‌های قبلی کاراتر خواهد بود. به خاطر وجود ابعاد بالای داده‌ها انجام این کار بسیار زمان‌گیر است. روش دوم برای انتخاب ویژگی فیلترینگ است که از ویژگی ذاتی داده‌ها استفاده می‌کند. ازجمله این ویژگی‌ها می‌توان به مواردی چون واریانس و شباهت ویژگی‌ها اشاره کرد (Alelyani, Tang, and Liu ۲۰۱۳, Dash and Koot ۲۰۰۹).

۲-۱-۲-۳-۲-۴- بهینه سازی فرا ابتکاری

امروزه بهینه‌سازی و محاسبات هوشمند یکی از حوزه‌های تحقیقاتی فعال است که کارهای تحقیقاتی زیادی بر روی آن در حال انجام است. در بسیاری از کارهای روزمره زمان، هزینه و منابع در اختیار ما محدود هستند، بنابراین استفاده بهینه از آن‌ها بسیار مهم هستند. در برنامه‌های مدرن، تغییر الگویی در طرز تفکر و طراحی در پیدا کردن راه‌حل‌هایی برای طرح‌های سبز و نگه‌دارنده انرژی انجام‌شده است. در دنیای واقعی مسائل راه‌حل بسیار سخت و غیر بدیهی دارند. برای مثال مسائل بهینه‌سازی ترکیبی که به دنبال یک راه‌حل بهینه از یکسری حالات محدود هستند، مانند مسئله فروشنده دوره‌گرد یک راه‌حل NP–سخت دارند و این بدان معنیست که یک راه‌حل بهینه به‌طور کل وجود ندارد.

روش‌های ابتکاری و فرا ابتکاری به‌طورکلی از طریق آزمون‌وخطا مسائل را حل می‌کنند؛ اما در روش‌های متاهیوریستیک از یک سطح بالاتری برای حل مسئله استفاده می‌کنند که در آن علاوه بر این روش از اطلاعات و گزینش برخی از راه‌حل‌ها برای هدایت فرایند جستجو استفاده می‌کند. الگوریتم‌های بهینه‌سازی روش‌ها و ابزارهایی هستند که برای حل مسائل بهینه‌سازی با هدف پیدا کردن بهینه مورداستفاده قرار می‌گیرند، هرچند این راه‌حل‌ها نمی‌توانند همیشه به چنین بهینه‌ای دست پیدا کنند. این امر بدان خاطر است که در دنیای واقعی همیشه عدم قطعیت وجود دارد. البته قابل‌توجه است به این نکته نیز اشاره کنیم که در دنیای واقعی یک سیستم بهینه همیشه قابل‌استفاده نیست، بلکه سیستمی قابل‌استفاده است که علاوه بر بهینه بودن قابلیت تحمل‌پذیری نیز در برابر خطا داشته باشد.

همان‌طور که بیان شد الگوریتم‌های بهینه‌سازی به دو دسته قطعی و غیرقطعی تقسیم می‌شوند. در الگوریتم‌های قطعی در هر بار اجرای الگوریتم ما به یک جواب ثابت و یکسانی می‌رسیم ولی در روش‌های غیرقطعی یا تصادفی هر اجراهای متفاوت الگوریتم به جواب‌های مختلفی می‌رسد که این به خاطر ذات تصادفی بودن این نوع الگوریتم‌ها است. به‌سادگی می‌توان برخی الگوریتم‌ها را از حالت قطعی به تصادفی و برعکس تبدیل کرد. برای مثال با تصادفی کردن نقطه شروع و باز شروع در الگوریتم بالا رفتن تپه الگوریتم از حالت قطعی به حالت غیرقطعی تبدیل می‌شود. به همین صورت می‌توان الگوریتم‌های قطعی را با تغییر اجزای آن با اجزای تصادفی، عدم قطعیت تکرارپذیری آن الگوریتم بالاتر می‌رود. چنین الگوریتم‌هایی را در اصطلاح ابتکاری و یا فرا ابتکاری گفته می‌شوند.

الگوریتم‌های فرا ابتکاری از طبیعت الهام گرفته شده‌اند و امروزه بسیار مورد کاربرد واقع شده‌اند. این الگوریتم‌ها مزیت‌های زیادی در میان الگوریتم‌های مختلف دارند. الگوریتم‌های فرا ابتکاری تنوع زیادی دارند، از الگوریتم ژنتیک گرفته تا تبرید تدریجی و کلونی مورچگان و زنبورعسل. از این الگوریتم‌ها در کاهش ویژگی نیز می‌توان استفاده کرد، به این صورت که تعداد ویژگی‌های متغیری (معیاری) است که می‌بایست نقطه بهینه آن‌ها را در بهینه کردن میزان دقت مدل به‌دست‌آمده از داده‌ها به دست آورد.

۲-۱-۳- معیارهای ارزیابی مدل‌سازها

برای ارزیابی یک سیستم بازیابی از پارامترهای زیر استفاده می‌شود:

$$precision = \frac{a}{a+b} \quad (۳-۲)$$

$$recall = \frac{a}{a+c} \quad (۴-۲)$$

$$F۱ = \frac{۲ * (precision * recall)}{precision + recall} \quad (۵-۲)$$

بدرستی بازیابی شده : a --- غیر مرتبط و اشتباه بازیابی شده: b --- مرتبط و بازیابی نشد: c

۲-۱-۴- نرمال سازی[1]

فرآیندی است که برای مقیاس بندی مجموعه‌ای از داده‌ها صورت می گیرد. در نرمال‌سازی خطی باید حاصل جمع تمام داده‌های خروجی (پس از نرمال‌سازی) برابر با ۱ شود. باید توجه داشت که نرمال‌سازی یک مجموعه‌داده با نرمال‌سازی یک بردار مشابه است اما یک تفاوت مهم بین آن‌ها وجود دارد. در نرمال‌سازی بردار، هر مؤلفه به جای تقسیم بر مجموع مؤلفه‌ها، بر مقدار بردار تقسیم می‌شود. یکی از کاربردهای نرمال‌سازی، مقایسه چندین داده با مقیاس متفاوت است (مانند مقایسه یک مجموعه‌داده بسیار بزرگ با یک مجموعه‌داده بسیار کوچک). در این وضعیت، نرمال‌سازی برای حذف تأثیر مقیاس‌های متفاوت به کار گرفته می‌شود.

نرمال‌سازی یا بی مقیاس سازی یک مفهوم زیربنایی در شیوه‌های تصمیم‌گیری چند معیاره نیز است. فرض کنید می‌خواهید دو نفر را بر اساس قد، وزن، سن و نمره مقایسه کنید. بدیهی است قد بر اساس سانتی‌متر، وزن بر اساس کیلوگرم و سن بر اساس سال است. هر معیار بر اساس مقیاس متفاوتی مورد سنجش قرارگرفته است بنابراین مقایسه این معیارها به سادگی میسر نیست .نرمال کردن در اینجا به معنای بی مقیاس کردن است.

۲-۱-۴-۱- نرمال کردن خطی

یک روش ساده برای نرمال کردن اعداد است که به محاسبه **بردار ویژه** نیز معروف شده است. در این روش کافی است هر عدد در یک مجموعه بر مجموع عناصر آن مجموعه تقسیم شود. در این صورت جمع کل عناصر پس از نرمال‌سازی یک خواهد بود.

$$n_{ij} = \frac{x_{ij}}{\sum_{1}^{m} x_{ij}} \qquad (۶-۲)$$

۲-۱-۴-۲- نرمال کردن برداری

[1] Normalization

از روش نرمال سازی برداری در شیوه‌فنی تاپسیس استفاده می‌شود که برای نرمال سازی مقادیر از روش برداری استفاده می‌شود. روش برداری برخلاف روش ساده نرمال سازی خطی به‌صورت زیر انجام می‌شود:

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{1}^{m} x_{ij}^{2}}} \qquad (۷-۲)$$

۲-۱-۴-۳- مقیاس‌گذاری ویژگی‌ها

ازآنجاکه محدود مقادیر داده‌های خام می‌تواند محدوده‌های وسیعی از داده‌ها را پوشش دهد، در برخی الگوریتم‌های یادگیری ماشین، برای هم مقیاس کردن ویژگی‌ها از این نوع روش‌ها استفاده می‌شود. در کاربردهای واقعی، به خاطر متفاوت بودن در انتخاب مقادیر ویژگی‌ها، یک ویژگی ممکن است اثر نامطلوبی بر ویژگی‌های دیگر بگذارد، بخصوص زمانی که از الگوریتم‌های خوشه‌بندی یا دسته‌بندی که از فاصله (اقلیدسی) استفاده می‌کنند استفاده می‌شود. نباید بدون اینکه نرمال‌سازی انجام شود از آن‌ها استفاده شود(Mohamad and Usman۲۰۱۳)؛ بنابراین نیاز است که بعدها را یکی شوند.

بنابراین برای صرفنظر کردن در انتخاب واحدهای اندازه‌گیری، داده‌ها باید استاندارد شوند. استانداردسازی تلاشی‌است تا همه متغیرها هم وزن شوند این امر زمانی مناسب است که هیچ اطلاعات قبلی از داده‌ها نداشته باشیم (Han, Pei, and Kamber۲۰۱۱).

مقیاس‌گذاری ویژگی‌ها روشی است که برای استاندارد کردن محدوده‌ای از متغیرهای مستقل یا ویژگی‌های داده‌ها انجام می‌شود. این عمل معمولاً در گام پیش‌پردازش انجام می‌شود و در برخی مواقع با نام نرمال‌سازی داده‌ها شناخته می‌شود. برخی روش‌های مقیاس‌گذاری ویژگی‌ها در ادامه آورده شده است.

۲-۱-٤-۳-۱- نرمال سازی حداقل حداکثر[۱]

[۱] Min-Max Normalization

یکی از ساده‌ترین روش‌های نرمال سازی است. این روش بر این مبنا است که محدوده اندازه‌های داده ویژگی‌ها را در بازه ۰ تا ۱ قرار دهد. فرمول این روش در زیر آورده شده است:

$$\bar{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (۸-۲)$$

مقدار $\bar{x}$ نرمال‌شده مقدار اصلی x است. این روش می‌تواند تحت تأثیر داده‌های پرت قرار گیرد و نتیجه دسته‌بندی یا خوشه‌بندی را تخت تاثیر قرار دهد.

۲-۱-٤-۳-۲- استانداردسازی[۱]

استاندارد سازی یا امتیاز-زد[۲]روش دیگری از مقیاس‌گذاری داده‌ها است که در بسیاری از روش‌های یادگیری ماشین (شبکه‌های عصبی، رگرسیون، ماشین بردار پشتیبان و ...) مورداستفاده قرار می‌گیرد. با این روش میانگین داده‌های هر ویژگی برابر صفر خواهد شد و واریانس آن‌ها برابر ۱ خواهد شد.

$$\bar{x} = \frac{x - average(data)}{Standard\_Variation(data)} \quad (۹-۲)$$

پژوهش (Mohamad and Usman ۲۰۱۳) نشان داد که استاندارد سازی در مقابل نرمال کردن حداقل_حداکثر باعث افزایش دقت بیشتری در خوشه‌بندی کا-میانه می‌شود. برای روش شدن این روش اگر چند دانشجو در چند آزمون متفاوت شرکت کنند و نمره‌های متفاوتی در مقیاس‌های متفاوتی بگیرند، آنگاه با نرمال‌سازی نمره‌ها می‌توان تشخیص داد که کدام دانشجو نمره بهتری را نسبت به بقیه دانشجوها کسب کرده و حتی چند درصد از دانشجوها کمتر از آن‌ها نمره را کسب کرده‌اند .

۲-۱-۵- انتخاب و استخراج کلیدواژه‌ها (نمایه‌سازی اسناد)

منظور از نمایه به صورت کلی، توصیفگر، اصطلاح نمایه‌ای، شناسه و کلیدواژه است. نمایه‌سازی در لغت نیز دارای معنای شاخص، نماد یا نشانه‌ای برای یک شی است و در اصطلاح انتساب نمادهای

---

[۱] Standardization

[۲] Z-Score

کلامی(کلیدواژه یا توصیفگر) و یا غیرکلامی(شماره رده بندی) به مدارک برای بازنمایی مفاهیم و محتوای آن‌ها است. یعنی اگر واژه‌ای که حکایت از محتوای یک مدرک دارد به آن مدرک نسبت داده شود نمایه‌سازی صورت گرفته است.

برای نمایه‌سازی نیازمند به یک زبان برای نمایه‌سازی هستیم. زبان نمایه‌سازی زبان موضوعی است. زبان موضوعی زبانی است که درآن واژه‌های واحد به طور طبیعی به کار می‌روند، اما واژه‌های‌این زبان شامل افعال نیستند. نمایه‌سازی با این زبان را نمایه‌سازی موضوعی می‌گویند که ممکن است مفهموم کامل آن سند بدون در نظر گرفتن فعل ناکافی باشد. پس زبان نمایه‌سازی به مجموعه واژگان مورد استفاده در نمایه‌سازی همراه با قواعد استفاده از آن را می‌گویند. درکل زبان‌های نمایه‌سازی را بر اساس معیارهای متفاوت به سه دسته متفاوت تقسیم کرده‌اند:

- نمایه‌سازی به زبان آزاد: که از کلمات داخل متن استفاده نمی‌شود و نمایه‌ساز از واژگانی که خود مناسب می‌داند استفاده می‌کند.
- نمایه‌سازی به زبان طبیعی: نمایه‌ساز از واژگان داخل متن استفاده می‌کند.
- نمایه‌سازی به زبان کنترل شده: که از واژگان استاندارد استفاده می‌کنند. این واژگان می‌توانند از اصطلاح نامه‌ها یا سرعنوان‌های موضوعی مورد استفاده قرار گیرند.
- از طرفی دیگر می‌توان نمایه سازی را بر اساس محتوای اسناد مورد تقسیم بندی زیر قرار داد:
- نمایه سازی واژه‌ای
- مبتنی بر عنوان: دراین نمایه‌سازی واژگانی که سند را توصیف می‌کنند از متن آن سند مورد استخراج قرار می‌گیرند. معمولا عنوان از اهمیت خاصی برخوردار است.
- مبتنی بر مآخذ: از منابع مورد استفاده در آن سند برای بدست آوردن حیطه آن سند مورد استفاده قرار می‌گیرد.
- نمایه سازی مفهومی :در نمایه سازی مفهومی؛ نمایه ساز بعد از خواندن متن سند، موضوع مرتبط با آن سند را به آن تخصیص می‌دهد. بر اساس‌اینکه این کلمات تخصیص یافته شده توسط نمایه ساز به چه صورتی ذخیر و بازیابی شوند به نظام‌های پیش هم آرا و پس هم آرا تقسیم می‌شود.

همانطور که گفته شد، در نمایه سازی به اسناد واژگانی تخصیص می‌دهیم که بتوانند آن سند را توصیف کنند. این واژگان می‌توانند از سند استخراج شوند یا با نظر خود نمایه ساز به سند داده شوند

و یا به صورت کنترل شده و با استفاده از اصطلاحنامه و یا کلمات استانداردی باشند. در هنگام تخصیص کلیدواژه‌ها به متن دو ویژگی اصلی را می‌بایست در نظر گرفت: میزان اهمیت کلمه در سند و قابلیت تفکیک پذیری سند توسط کلمه در مجموعه اسناد. قابلیت تفکیک‌پذیری سند به شکل‌های مختلف قابل ارزیابی است که در (Wartena, Brussee, and Slakhorst ۲۰۱۰) روشی برای میزان تفکیک‌پذیری کلمات بر اساس توزیع هم‌رخدادی کلمات در متن پیشنهاد کرده است.

برای استخراج و تخصیص کلمات کلیدی می‌توان دو معیار را در نظر گرفت:

- استخراج عبارات تک کلمه‌ای[۱]
- استخراج عبارات چند کلمه‌ای[۲]

هرکدام ازاین معیارها ویژگی‌های مخصوص به خودشان را دارند، مثلا اگر کلمات تک مدنظر گرفته شوند و چندکلمات حذف شوند در بازیابی اطلاعات با مشکل مواجه می‌شویم و کلماتی که با LC توصیف شوند در بازیابی کافی نیستند. بهترین راه حل استفاده از LC و SC به طور همزمان هستند (Okada et al. ۲۰۰۱). حال پیدا کردن این عبارات می‌توانند به گونه‌های متفاوتی انجام شود. دو نوع روش برای پیدا کردن کلمات کلیدی در (Abilhoa and de Castro ۲۰۱۴) ارائه‌شده است:

۱- **بر اساس یک مجموعه اسناد:** دراین روش، سیستم با مطالعه تمام اسناد حدس می‌زند که چه کلمات کلیدی در آن‌ها وجود دارد.

۲- **بر اساس یک سند:** سیستم با گرفتن اطلاعات آماری و زبان شناختی تنها یک سند و بررسی آن کلمات کلیدی از آن استخراج می‌کند. در روشی که در (Matsuo and Ishizuka ۲۰۰۴) برای استخراج کلمه کلیدی از یک سند ارائه‌شده به صورتی است که ابتدا تعدادی کلمات پرتکرار آن سند پیدا شده و سپس میزان هم رخدادی این کلمات با دیگر کلمات متن محاسبه می‌شود و در صورتی که این هم رخدادی زیاد باشد، آن کلمات به عنوان

[۱] Single Component (SC)

[۲] Long Component (LC)

کلمات با اهمیت در نظر گرفته می‌شوند. درجه هم رخدادی کلمات با پارامتر χ^2 اندازه گرفته شده است.

همانطور که قبلا گفته شد کلمات کلیدی می‌توانند هم از متن استخراج شوند هم بعد از مطالعه سند حدس زده شوند. روش‌هایی که کلمات کلیدی را از متن استخراج می‌کنند از ویژگی‌های زیر بهره می‌گیرند (Oelze۲۰۰۹)

۱. **شیوه‌های آماری ساده:** دراین روش اطلاعات آماری ساده‌ای را مورد استخراج قرار می‌دهند که این اطلاعات شامل موارد زیر است:

- فرکانس کلمه[۱]
- *معکوس فرکانس سند*[۲]
- ضرب فرکانس کلمه در معکوس فرکانس سند
- محل رخداد کلمات
- همرخدادی کلمات (ارتباطات گرافی (Liu et al.۲۰۰۹) (Rose et al.۲۰۱۰))؛
- موقعیت وقوع کلمات (اولین رویداد، آخرین رویداد و ...)
- استفاده از ارتباطات گرافی

۲. شیوه‌های زبان شناختی[۳] (Hulth۲۰۰۳) و (Hong and Zhen۲۰۱۲)

- استفاده از جز سخن[۴]
- ساختار نحوی[۵]
- کیفیت معنایی[۶]

---

[۱] Term Frequency

[۲] Inverse Document Frequency

[۳] linguistic

[۴] Part Of Speech (POS)

[۵] Syntactic Structure

[۶] Semantic Qualities

- زنجیره لغات[1]

۳. یادگیری ماشین (Mihalcea and Tarau ۲۰۰۴)

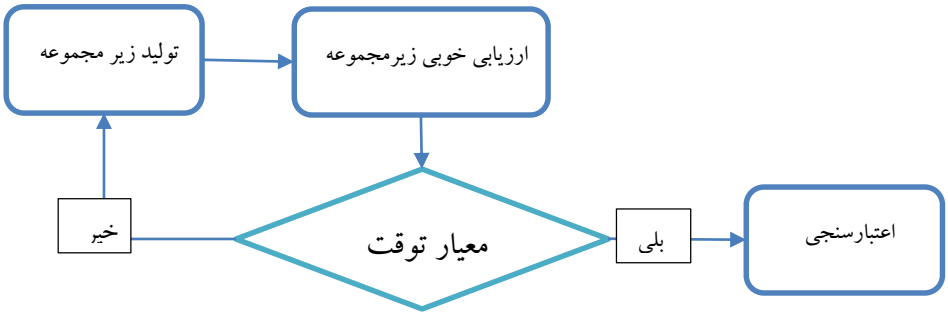
- شبکه بیز
- ماشین بردار پشتیبان
- KEA

۴. ترکیبی (مانند (Ercan and Cicekli ۲۰۰۷))

به طور کلی می‌توانیم روش‌های انتخاب‌ویژگی را برای انتخاب کلیدواژه‌های مورد نیاز استفاده کنیم. روش‌های انتخاب ویژگی به طور کلی برای انتخاب مراحل زیر مورد استفاده قرار می‌گیرد:

- در گام اول یک زیرمجموعه ای از ویژگی‌های اولیه که داریم انتخاب می‌کنیم. این بخش به نام تولید می‌تواند نام گذاری شود که از روش‌های زیر می‌تواند استفاده کند:
  - استفاده از یک مجموعه خالی و اضافه کردن ویژگی‌های نامطلوب در هرگام (روش رو به جلو)
  - استفاده از همه مجموعه‌ها و حذف ویژگی‌های نامطلوب در هر گام (رو به عقب)
  - استفاده از یک مجموعه تصادفی و اضافه و کم کردن یک و یا چند ویژگی در هر گام.
- در گام بعدی این زیرمجموعه را به تابعی برای ارزیابی میزان بهینه بودن و کارایی مورد تحلیل قرار خواهیم داد.
- خروجی گام قبل را با معیارهای توقف مورد سنجش قرار می‌دهیم تا در صورتی که مورد تایید قرار گرفت به بخش اعتبار سنجی و پایانی برود ولی در صورتی که مورد تایید قرار نگرفت به گام اول رفته و یک زیر مجموعه جدیدی از ویژگی‌ها انتخاب خواهند شد.

[1] Lexical chain



شکل ۲-٦ مراحل انتخاب ویژگی

معیار پایان برای برنامه می‌تواند معیارهای گوناگونی را از جمله موارد زیر شامل شود:

- مقدار ویژگی‌های مقصد از قبل مشخص شده باشد و به آن مقدار رسیده باشیم.
- تعداد مراحل طی شده از قبل مشخص باشد.
- تغییر در تعداد ویژگی‌ها، بهبودی نتیجه‌ای حاصل نکند.
- براساس معیار ارزیابی یک زیر مجموعه بهینه بدست آید.

۲-۱-۶- قواعد انجمی یا قواعد وابستگی

یادگیری قواعد انجمنی یکی از روش‌های یادگیری مبتنی بر قانون است که برای کشف ارتباط بین متغیرها در پایگاه داده‌های بزرگ مورداستفاده قرار می‌گیرد. هدف این روش‌ها کشف قوانین محکم با استفاده از معیارهای مورد علاقه کاربر است. با استفاده از قوانین استخراج‌شده می‌توان داده‌های بیشتری را تحلیل کنیم.

جست‌وجو به دنبال روابط در پایگاه داده‌های بزرگ با عنوان تحلیل وابستگی یا کاوش قواعد وابستگی[۱] شناخته شده است. مسئله آن است که کشف ترکیب‌های گوناگون از اقلام می‌تواند کاری زمان‌بر و چه‌بسا از لحاظ توان پردازشی غیرممکن باشد. این ترکیب‌ها به دو شکل ممکن است به وقوع بپیوندند. این دو فرم عبارت‌اند از اقلام مکرر و قواعد وابستگی. مجموعه اقلام مکرر مجموعه‌ای از آیتم‌ها هستند که مکرراً با یکدیگر به وقوع می‌پیوندند. دومین راه برای مشاهده روابط جالب بین

[۱] Association rule mining

آیتم‌ها، قواعد وابستگی است. قواعد وابستگی، ارتباطات قدرتمند بین آیتم‌ها را تبیین می‌کنند. «کاوش قواعد وابستگی» برای داده‌های دسته‌ای مناسب است.

قوانین انجمنی روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند. پیدا کردن چنین قوانینی می‌تواند در حوزه‌های مختلف موردتوجه بوده و کاربردهای متفاوتی داشته باشد به‌عنوان مثال کشف روابط انجمنی بین حجم عظیم تراکنش‌های کسب‌وکار می‌تواند در تشخیص تقلب، در حوزه پزشکی و همچنین داده‌کاوی در مورد اطلاعات روش به‌کارگیری وب توسط کاربران و شخصی‌سازی مورداستفاده قرار گیرد یا در طراحی کاتالوگ، بازاریابی و دیگر مراحل فرایند تصمیم‌گیری کسب‌وکار مؤثر باشد. الگوریتم آپریوری یکی از بهترین الگوریتم برای کاوش قوانین وابستگی است. این الگوریتم از استراتژی جستجوی اول-سطح برای شمارش پشتیبان مجموعه آیتم‌ها استفاده می‌کند و با استفاده از یک تابع تولید کاندید، از خصوصیت بستار رو به پایین پشتیبان بهره می‌برد.

یک مثال از کاربردهای قواعد وابستگی، موتورهای جست‌وجو است. برای کاوش مجموعه اقلام مکرر و کشف قواعد وابستگی، از پشتیبان و اطمینان به‌عنوان راهکارهایی استفاده می‌شود که با بهره‌گیری از آن‌ها می‌توان تحلیل وابستگی را کمّی کرد. پشتیبان یک مجموعه اقلام، به‌عنوان درصدی از مجموعه‌داده که شامل آن مجموعه اقلام تعریف‌شده است. اطمینان برای قاعده $P \rightarrow H$ به‌صورت $support(P \mid H) / support(P)$ تعریف شده است. در پایتون، علامت «|» برای اتحاد مجموعه‌ها به کار می‌رود، این در حالی است که نماد ریاضی اتحاد مجموعه‌ها «U» است. $P \mid H$ یعنی (اتحاد P و H) شامل همه آیتم‌های موجود در مجموعه P یا H است.

تعاریف زیر در محاسبه قواعد انجمنی مورداستفاده قرار می‌گیرد:

- **پشتیبان**[1]: نشان می‌دهد که یک مجموعه اقلام خاص و موردنظر چند بار در پایگاه تکرار شده است. این مقدار به‌عنوان کسر رکوردهای شامل XUY بر کل تعداد رکوردها در پایگاه داده است. به‌عنوان نمونه درصورتی‌که پشتیبان یک نمونه (الگو یا مجموعه اقلام) ۱۰ درصد

[1] Support

باشد، در این صورت تنها ۱۰ درصد از تراکنش‌ها حاوی آن مورد هستند.

Support (XY) = Support count of (XY) / Total number of transaction in D

- **اطمینان**[1]: حاصل تقسیم تعداد تراکنش‌های حاوی XUY به کل تعداد رکوردهای شامل x است. این مقدار معیاری برای نشان دادن میزان استحکام قواعد وابستگی است. به‌عنوان نمونه درصورتی‌که اطمینان یک قاعده وابستگی مثل $X \Rightarrow Y$ برابر ۹۰ درصد باشد در این صورت باشد، رد این صورت ۹۰ درصد از تراکنش‌هایی که حاوی X هستند، شامل Y نیز می‌شوند.

Confidence (X|Y) = Support (XY) / Support (X)

(۲۷-۲)

پیدا کردن مجموعه اقلام مکرر

یک روش برای پیدا کردن مجموعه اقلام مکرر الگوریتم اپریوری است. این الگوریتم نیازمند آن است که حداقل میزان اطمینان به‌عنوان ورودی به آن داده شود. الگوریتم، لیستی از همه اقلام کاندید با یک آیتم را فراهم می‌کند. سپس، مجموعه‌داده تراکنشی به‌منظور بررسی آنکه کدام مجموعه‌ها دارای حداقل سطح پشتیبان هستند اسکن می‌شود. مجموعه‌هایی که حداقل سطح پشتیبان را ندارند، حذف می‌شوند. مجموعه‌داده‌های باقیمانده ترکیب می‌شوند تا مجموعه اقلام دارای دو آیتم حاصل شوند. دوباره، مجموعه‌داده تراکنشی اسکن می‌شود و مجموعه اقلامی که دارای حداقل سطح پشتیبانی نیستند، بیرون انداخته می‌شوند. این روال تا زمانی که همه مجموعه‌ها حذف شوند ادامه می‌یابد. برای کشف قواعد وابستگی، کار با یک مجموعه اقلام مکرر آغاز می‌شود. داده‌کاوی می‌داند که این مجموعه اقلام یکتا هستند، اما قصد دارد بررسی کند که آیا می‌توان نکته دیگری از آن استخراج کرد. یک آیتم یا یک مجموعه از اقلام می‌تواند دلالت ضمنی بر دیگر آیتم داشته باشد.

[1] Confident

۲-۱-۷- بازنمایی کلمه یا تعبیه‌کردن کلمات[۱]

بسیاری از الگوریتم‌های یادگیری ماشین و اکثر روش‌های یادگیری عمیق توانایی پردازش رشته‌ها یا متن‌های ساده را به‌صورت خام ندارند. این روش‌های نیازمند استفاده از عددها برای نمایش کلمات به‌عنوان ورودی برای هرگونه روشی مانند دسته‌بندی، خوشه‌بندی یا رگرسیون هستند. بنابراین باید کلمات به صورتی برای رایانه مشخص شود که سیستم بتواند فرق بین سیب خوراکی را با شرکت سیب متوجه شود و یا اینکه در هنگام جستجو لغت «مسی» بتوان «رونالدو» و یا «فوتبال» را بازیابی کرد.

بنابراین به دلیل در اختیار داشتن حجم عظیمی از داده‌ها، استخراج دانش از آن‌ها و ساختن برنامه‌های کاربردی امری بسیار ضروری است. برخی از کاربردهای برنامه‌های متنی عبارت‌اند از: تحلیل رفتار و احساسات مخاطب و دسته‌بندی یا خوشه‌بندی اخبار توسط شرکت گوگل است.

نمایش یک کلمه در پردازش زبان طبیعی به‌صورت نمایش یک شی ریاضی برای هر کلمه است که این نمایش عموماً به‌صورت برداری است. هر عضو از بردار نیز یک ویژگی از آن کلمه (مثلاً رابطه معنایی یا تفسیر گرامری از آن کلمه) محسوب می‌شود (Turian, Ratinov, and Bengio ۲۰۱۰).

بازنمایی کلمات اولین بار در سال ۱۹۸۶ در (Rumelhart, Hinton, and Williams۱۹۸۶) معرفی شد. می‌توان بازنمایی کلمات یا ***بازنمایی کلمات*** را به‌صورت زیر تعریف کرد: « بازنمایی کلمات به‌صورت کلی سعی می‌کند تا هر کلمه را با استفاده از یک دیکشنری (مجموعه‌ای از کلمات) به یک بردار نگاشت کند». به‌عبارت‌دیگر بازنمایی کلمات تبدیل کردن کلمات به برداری از اعداد است (یا به عبارتی نمایش برداری عددی کلمات). یک بردار نمایش کلمه می‌تواند برداری باشد که عدد یک نماینده موقعیت وجود کلمه، و عدد صفر نمایانگر عدم وجود کلمه است. بنابراین کلمات در نمایش برداری دارای ابعاد بسیار زیادی هستند. در بازنمایی کلمات بردار نمایش کلمات شبیه، نزدیک هم هستند.

[۱] Word embedding

همه روش‌های بازنمایی کلمات بر این اصل استوار هستند که کلماتی که در یک محتوا آمده است معانی مشابهی دارند (Firth۱۹۵۷). در روش‌های مختلف بازنمایی کلمات ازآنجاکه کلمات در محتواهای مختلف می‌توانند معانی مختلفی داشته باشند، برای کلمات در محتواهای مختلف بردارها و نمایش‌های متفاوتی ارائه می‌دهند. برای مثال «لیو» در سال ۲۰۱۵ در مقاله (Liu et al. ۲۰۱۵) سه مدل برای نمایش کلمات در محتواهای متفاوت ارائه کرده است. برای ارزیابی این سه روش ازنظر شباهت معنایی کلمات و خوشه‌بندی استفاده‌شده است. بازنمایی کلمات‌ها انواع مختلی دارند که در زیر به چند مورد آن اشاره شده است:

- بازنمایی کلمات مبتنی بر تکرار
  - بردار شمارش
  - بردار TF-IDF
  - بردار هم‌رخدادی
- بازنمایی کلمات مبتنی بر پیش‌بینی
  - روش Vec۲Word

۲-۱-۷-۱- روش‌های ارزیابی بازنمایی کلمات

طبق مطالعات انجام‌شده استاندارد خاصی برای ارزیابی روش‌های بازنمایی کلمات نیست. روش‌های زیر برای ارزیابی روش‌های بازنمایی کلمات در مقالات مختلف مورداستفاده قرار گرفته‌اند:

- مقایسه‌ای[۱] در این مقایسه مجموعه‌داده‌ای درست می‌شود که شامل دو کلمه است و بعد از آموزش آزموده می‌شود که کلمه اول چقدر قابلیت پیش‌بینی کلمه دوم را دارد.
- مدل‌ها: در این روش کلمات به‌عنوان ورودی به یک مدل یادگیری (مانند خوشه‌بندی یا دسته‌بندی) اعمال می‌شوند.
  - دسته‌بندی با ویژگی‌های جدید و ارزیابی آن توسط پارامترهای ارزیابی.

[۱] Analogy

- خوشه‌بندی با ویژگی‌های جدید و ارزیابی آن توسط پارامترهای ارزیابی.

- *مجموعه‌داده*‌های شباهت[1] این *مجموعه‌داده*‌ها بسیار کلی هستند و یک دامنه خاصی را شامل نمی‌شوند.
- *مجموعه‌داده*‌های شخصی

۲-۱-۷-۲- روش «کلمه-به-بردار[2]»

بازنمایی کلمات تاریخچه طولانی در تحقیقات علمی دارد، برای مثال در سال ۲۰۰۳ (Bengio et al.۲۰۰۳) با استفاده از شبکه عصبی مدل زبانی آماری و بازنمایی کلمات را آموزش داده است. بازنمایی کلمات تا سال ۲۰۱۳ که «میکولوو» دو مدل ساده با کارایی بسیار بالاتر از روش‌های قبل ارائه نداده بود، جهش زیادی نکرد. روش‌های ارائه‌شده توسط «میکولوو» باعث کاهش بسیار زیاد در زمان محاسبه نیز شده است (Mikolov et al.۲۰۱۳). هرچند که این مدل برای یادگیری نیاز به مجموعه اسناد زیادی دارد اما می‌تواند یک زیربنای مهم و قابل‌اطمینانی برای دیگر پردازش‌های زبانی باشد. البته لازم به ذکر است ازآنجاکه ۲۰٪ از مفهوم کلمات به ترتیب کلمات ارتباط دارد روش vec۲word بخشی از مفاهیم را از دست می‌دهد (Lai et al.۲۰۱۶).

مبنای روش Vec۲Word بر این اصل استوار است که کلماتی که در یک متن وجود دارند معنای مشابهی نیز دارند (Harris۱۹۵۴). روش کار Vec۲Word به این صورت است که با بردارهایی که به‌صورت تصادفی مقداردهی می‌شوند شروع به کار می‌کند. متن را به‌صورت ترتیبی مورد اسکن قرار می‌دهد و معمولاً با پنجره‌ای ثابت اطراف کلمه موردنظر را موردبررسی قرار می‌دهد. الگوریتم ضرب نقطه‌ای بین کلمه هدف و کلمه‌های محتوا انجام می‌دهد و سعی بر حداقل رساندن SGD[3] دارد. هر وقت که دو کلمه در یک متن مشابه قرار داشته باشند، ارتباط یا فاصله فضایی آن‌ها قوی‌تر می‌شود. در ابتدا برای روش Vec۲Word استفاده از یکسان‌ساز Softmax سلسله‌مراتبی پیشنهاد شد

[1] Similarity

[2] Word۲Vec (W۲V)

[3] Stochastic Gradient Descent

و در ادامه روش جدیدی به اسم نمونه‌برداری منفی[۱] ارائه گردید. این روش بسیار ساده‌تر و مؤثر بوده است. فرضیه اولیه این روش این است که زمانی که فاصله بین بردارها حداقل می‌شوند، چند کلمه به‌صورت تصادفی انتخاب‌شده و فاصله بین آن‌ها را تا بردار هدف حداکثر می‌شود. با این عمل کلمات نامشابه از یکدیگر دور می‌شوند.

روش Vec۲Word نشان داد که یک روش کارا و کارآمد در تشکیل بازنمایی کلمات در مجموعه‌داده‌های ساختار نیافته در ابعاد بزرگ است. معماری این روش شامل دو مدل است: مدل «بسته کلمات پیوسته»[۲] و مدل «اسکیپ گرام»[۳] (Mikolov et al. ۲۰۱۳). به لحاظ الگوریتمی این دو مدل بیان‌شده مشابه هم هستند با این تفاوت که «بسته کلمات پیوسته» لغات هدف را از روی لغات متن ورودی پیش‌بینی می‌کند اما اسکیپ گرام از روی لغات هدف لغات ورودی را پیش‌بینی می‌کند.

روش CBOW می‌تواند روشی مفید برای استفاده در مجموعه دادگان کوچک‌تر باشد. درصورتی‌که قصد خوشه‌بندی دامنه محدودی از متون (مانند چکیده‌های مقالات) را داشته باشیم به خاطر وجود تعداد کم متن و تکرار کم کلمات می‌توانند نتیجه خوشه‌بندی را کم دقت کرده و با تداخل کلمات *تشخیص زیردامنه‌ها (موضوع اصلی) سخت بشود* (Li et al. ۲۰۱۸).

در مقایسه‌ای که در عملکرد LSA با vec۲word انجام شد، نتایج نشان داد که در مجموعه‌داده‌هایی با اندازه ۵میلیون کلمه بیشتر، روش vec۲word از LSA بهتر عمل می‌کند (Altszyler et al. ۲۰۱۶). برای تشکیل یک بازنمایی کلمه مناسب (با استفاده از vec۲word و دیگر روش‌ها) در پژوهش (Lai et al. ۲۰۱۶) نشان داده‌شده است که انتخاب یک دامنه مرتبط مهم‌تر از اندازه مجموعه متون است.

روشی که در (Li et al. ۲۰۱۸) برای بهبود در کیفیت روش‌های بازنمایی کلمات در متون کوتاه (مانند چکیده مقالات علمی) پیشنهادشده است به این قرار است که: ابتدا انتخاب بخش اول چکیده (چکیده به سه قسمت: هدف، روش، و جمع‌بندی تقسیم‌شده) که مرتبط با موضوع مقاله و متن است

---

[۱] Negative Sampling

[۲] Continous Bag-of-Word (CBOW)

[۳] Skip-gram

انجام می‌شود، سپس حذف کلمات پرتکراری که در پردازش‌های قبلی حذف نشده‌اند (لیست کلمات‌ایست جدید) صورت می‌پذیرد و در ادامه از کلمات *به‌دست‌آمده* در آموزش خوشه‌بندی و استخراج خوشه‌های مناسب مورداستفاده قرار گرفته‌اند. برخی پارامترهای روش Vec۲Word به‌قرار زیر است:

- **زیر-نمونه‌برداری:** برای حذف کلماتی که به‌صورت پرتکرار در تمام اسناد وجود دارند و با اکثر کلمات *هم‌رخداد* هستند مورداستفاده قرار می‌گیرد.
- **حداقل تکرار:** برای کنترل کردن دایره لغات کلماتی که از یک آستانه تعداد تکرار بیشتر در متن‌ها نیامده‌اند از دایره لغات حذف خواهند شد.
- **نرخ یادگیری:** برای آموزش شبکه‌های عصبی نرخ یادگیری لحاظ می‌شود.
- **اندازه بردار:** اندازه بردار منظور اندازه برداری است که کلمه با آن بعد از آموزش نشان داده می‌شود. هرچقدر اندازه بردار بیشتر باشد نمایش کلمه نیز بهتر خواهد بود.
- **اندازه پنجره محتوا:** اندازه پنجره محدوده کلماتی را در متن مشخص می‌کند که در مجاورت کلمه هدف قرار دارند. برای مثال اندازه ۵ بیان می‌کند که ۵کلمه قبل و ۵ کلمه بعد از کلمه هدف در آموزش لحاظ خواهند شد.
- **اندازه نمونه‌برداری منفی (neg).**

در بخش یادگیری vec۲word پژوهش (Lai et al. ۲۰۱۶) و (Chiu et al. ۲۰۱۶) نشان داده است که اندازه بردار ۵۰ برای پردازش‌های زبان طبیعی کافی است. اما پژوهش (Buscaino ۲۰۱۴) نشان می‌دهد که برای *به دست* آوردن دقت کافی طول بردار ۱۰۰ کافی است و طول بردار بزرگ‌تر از ۱۰۰ افزایشی در مقدار دقت ندارد (پژوهش برای تعداد چکیده‌های مختلف (از ۳۰ هزار تا ۱۵۰ هزار چکیده) انجام‌شده است). از طرفی در (Buscaino ۲۰۱۴) نشان می‌دهد که برای چکیده‌های بزرگ‌تر شاید طول بردار کوچک‌تر کفایت کند ولی برای تعداد چکیده‌های کمتر طول بردار ۵۰ دقتی در حدود ۶۰ تا ۷۰ درصد و با طول بردار ۱۰۰ دقت به حدود ۹۰ درصد خواهد رسید.

روش vec۲word کاربردهای بسیاری می‌تواند داشته باشد که ازجمله کاربردهایش می‌توان به موارد زیر اشاره کرد: استفاده در تشخیص استخراج موضوع از داده‌های کتابشناختی (Zhang et al.

۲۰۱۸)، تشکیل سیستم بازیابی اطلاعات (Nikitinsky, Ustalov, and Shashev ۲۰۱۵)، دسته‌بندی احساسات نظرات (Zhang et al. ۲۰۱۵)، ارزیابی معنایی سامانه‌های خلاصه‌سازی متن (Campr and Ježek ۲۰۱۵)، مقایسه تشابه بین دو سند(Kusner et al. ۲۰۱۵)، دسته‌بندی احساسات در توییتر(Tang et al. ۲۰۱۴)، دسته‌بندی اسناد (Lilleberg, Zhu, and Zhang ۲۰۱۵) ، کاهش ویژگی در پردازش داده‌های عظیم (Ma and Zhang ۲۰۱۵) و ترکیب آن برای استخراج ویژگی اسناد (برای دسته‌بندی) با ترکیب با روش LDA (Wang, Ma, and Zhang ۲۰۱۶)

توجه شود ازآنجاکه بردار کلمات در روش جایگذاری کلمات نشان‌دهنده معنای کلمه است، و ازآنجاکه بردار هر کلمه بر اساس وقوع و با استفاده از کلمات مجاورش ساخته می‌شود، کلمات دارای تکرار پایین ازنظر معنایی دارای خطا می‌توانند باشند (Li and Wang ۲۰۱۷). از طرفی دیگر در این فضا کلماتی که ازنظر معنایی مشابه هم هستند در یک خوشه قرار می‌گیرند و یک گروهی از کلمات را تشکیل می‌دهند (Wang et al. ۲۰۱۵).

روش‌هایی نیز برای بهبود عملکرد vec۲word (برای مثال در حالت‌هایی که حجم کلمات کم باشد و یا دانشی در خصوص سند نباشد) در (Bian, Gao, and Liu ۲۰۱۴) و (Li and Wang ۲۰۱۷) ارائه‌شده است. با استفاده ترکیب روش Vec۲Word با LDA (Li et al. ۲۰۱۸) و حذف نویز در پایگاه‌داده‌های کوچک‌تر (یک‌سوم قبلی‌ها) بهبود حاصل‌شده است (Mnih and Kavukcuoglu ۲۰۱۳). یکی دیگر از معایب vec۲word مشخص نشدن کلمات مهم با این روش است. در برخی از پژوهش‌ها برای رفع این مشکل، از روش‌هایی مانند استفاده از tf*idf استفاده‌شده است (Lilleberg, Zhu, and Zhang ۲۰۱۵).

۲-۱-۷-۳- هم‌رخدادی کلمات

هم‌رخدادی کلمات یک‌راه کلاسیک برای نمایش کلمات به‌صورت برداری است. در این بردار کلمات را بر اساس میزان هم‌رخدادی آن‌ها با یکدیگر در بردار نشان داده می‌شوند. هر عضو بردار نماینده یک کلمه است. بنابراین عددی که در آن خانه از بردار قرار داده می‌شود میزان هم‌رخدادی با آن کلمه را نشان می‌دهد. ازآنجاکه هر کلمه با تمام کلمات دیگر *هم‌رخداد* نیست، روش

هم‌رخدادی کلمات از کم‌پشتی بردارها رنج می‌برد (Mnih and Kavukcuoglu ۲۰۱۳). همردخدادی کلمات در پیدا کردن موضوع متن مورداستفاده قرار می‌گیرد برای مثال در (de la Hoz-Correa, Muñoz-Leiva, and Bakucz ۲۰۱۸) بدین منظور استفاده‌شده است.

راه‌کارهای هم‌رخدادی روش‌های سنتی برای استخراج عنوان سند از مجموع اسناد هستند. در مواقعی که اندازه و تعداد اسناد کم باشند کم‌پشت بودن هم‌رخدادی کلمات مشکلی است که در این‌گونه روش‌ها به وجود می‌آید. در برخی از منابع تحلیل هم‌رخدادی کلمات شامل پنج گام اصلی است: استخراج داده، استاندارد کردن کلیدواژه‌ها، ساخت ماتریس هم‌رخدادی، خوشه‌بندی کلمات، نمایش گروه‌های کلمات (Dehdarirad, Villarroya, and Barrios ۲۰۱۴).

بر این اساس که هم‌رخدادی کلمات محتوای سندهای یک فایل را مورد توصیف قرار می‌دهند. با استفاده از هم‌رخدادی کلمات در نقشه‌های کتاب‌سنجی، با ترسیم زیر حوزه‌ها، ارتباط بین آن‌ها و تشخیص روند آینده آن‌ها، می‌توان آینده را مورد پیش‌بینی قرار دهیم (Callon, Courtial, and Laville ۱۹۹۱).

روش‌هایی برای بازنمایی اسناد نیز وجود دارد که در یکی از آن‌ها برای نمایش بهتر و با مفهوم بیشتر یک سند «ونگ» در سال ۲۰۱۶ در (Wang, Ma, and Zhang ۲۰۱۶) ارائه کرده است. در پژوهش وی روشی برای استخراج ویژگی اسناد به‌کاربرده است. در این روش با استفاده از vec۲word و ترکیب آن با روش LDA نمایشی بهتر برای اسناد ارائه شده است. ازآنجاکه هر سند می‌تواند چند موضوع از LDA را داشته باشد برای هر موضوع (که شامل چند کلمه است) یک بردار تشکیل می‌شود. برای نشان دادن عملکرد این روش از دسته‌بندی استفاده‌شده است.

۲-۱-۸- خوشه‌بندی اسناد

تحلیل خوشه‌بندی هنر پیدا کردن گروه‌هایی در داخل داده‌ها است که اعضای آن دارای ویژگی‌هایی مشابه یکدیگر باشند (Kaufman and Rousseeuw ۲۰۰۹). تحلیل خوشه‌بندی استفاده و اعمال یک و یا چند الگوریتم خوشه‌بندی باهدف پیدا کردن الگوها یا دسته‌بندی یک مجموعه‌داده است. در الگوریتم‌های خوشه‌بندی داده‌های داخل یک خوشه نسبت به داده‌های خوشه‌های دیگر،

شباهت بیشتری نسبت به یکدیگر دارند. معیارهای شباهتی که با آن خوشه‌ها را می‌توان مدل کرد شامل: فاصله اقلیدسی، فاصله احتمالی، و یا دیگر انواع فاصله‌ها است. بنابراین تحلیل خوشه یک ابزار اکتشافی پایه‌ای است که زمانی که عضویت داده‌ها در گروه‌ها مشخص نیست، بردارهای داده‌ها را در گروه‌هایی که مشابه هم هستند مرتب می‌کند. تفاوت روش‌های خوشه‌بندی از استفاده از معیارهای شباهت و فاصله بین بردارهای داده‌ها نشاءت می‌گیرد (Wilks۲۰۱۱).

تحلیل خوشه‌ای، گروه و یا خوشه‌ای از داده‌ها را درست می‌کند که این خوشه‌ها به‌گونه‌ای تشکیل می‌دهد که اشیاء داخل یک خوشه بسیار به هم شبیه و اشیاء خوشه‌های مختلف بسیار از هم دور هستند. معیار شباهت وابسته به کاربرد و هدف خوشه‌بندی دارد. تحلیل خوشه‌بندی یک روش یادگیری غیرنظارتی است و یک کار بااهمیت در تحلیل داده‌های اکتشافی است. از ویژگی‌های بارز هریک از این الگوریتم‌ها معیار اندازه‌گیری شباهت آن‌ها است. بر اساس (Han, Pei, and Kamber ۲۰۱۱)دسته‌بندی الگوریتم‌های مهم خوشه بند به‌قرار زیر است:

- **خوشه‌بندی سلسله مراتبی:** یکی از روش‌های خوشه‌بندی است که بر مبنای ایده اصلی فاصله اشیای نزدیک، نسبت به اشیاء دورتر ارائه‌شده است. در فاصله‌های مختلف، خوشه‌های متفاوتی شکل می‌گیرند که با ساختن سلسله مراتبی از خوشه‌ها به شکل درخت و یا دندوگرام می‌توانند نشان داده شوند. در این روش درخت تنها یک مجموعه خوشه ساده نیست، بلکه یک سلسله‌مراتب چند سطحی است که ترکیب چند خوشه در یک سطح به‌عنوان خوشه‌ای در یک سطح بالاتر در نظر گرفته می‌شوند. این امر به ما این اجازه را می‌دهد که سطح خوشه‌بندی خودمان را به صورتی تعیین کنیم که بیشترین تناسب را برای کاربرد موردنظر داشته باشد. خوشه‌بندی سلسله مراتبی به دودسته تک‌پیوندی[۱] (پایین به بالا یا متراکم شدن[۲]) و پیوند کامل[۳] (بالا به پایین یا تقسیم شونده[۴]) تقسیم می‌شود.

[۱] Single linkage

[۲] Agglomerative

[۳] complete linkage

[۴] Divisive

- **روش‌های تقسیم‌بندی**[1]: در این‌گونه روش‌ها مجموعه‌داده‌ها به تعداد مشخصی قسمت (که خوشه گفته می‌شود) تقسیم می‌شوند. این خوشه‌ها به صورتی تشکیل می‌شوند که معیار (که در اکثر مواقع معیار شباهت گفته می‌شوند) تقسیم‌بندی را به‌خوبی بهینه کنند. خوشه‌بندی کا-میانه[2] از معروف‌ترین این روش‌ها است که یکی از روش‌های خوشه‌بندی مرکزی[3] است که در این نوع روش مرکز خوشه‌ها با یک بردار مرکزی نشان داده می‌شوند، که ممکن است لزوماً جزء مجموعه‌داده نباشد. در روش کا-میانه داده‌ها را به k قسمت مجزا (منحصربه‌فرد) ، مبتنی بر فاصله تا مرکز ثقل خوشه تقسیم می‌کند. در این روش در آغاز به تعداد مشخص نقاطی به‌عنوان مرکز خوشه (که در ابتدا به‌صورت تصادفی تولید می‌شوند) در نظر گرفته می‌شود. سپس با بررسی هر داده، آن را به نزدیک‌ترین مرکز خوشه نسبت می‌دهیم.) ازجمله مشکلات این روش این است که بهینگی آن وابسته به انتخاب اولیه مراکز بوده و بنابراین بهینه محلی است. این روش مناسب‌تر برای حجم زیادی از داده‌ها هستند اما برای حجم بسیار زیاد داده مناسب نیست. مشکل دیگر آن تعیین تعداد خوشه‌ها در آغاز خوشه‌بندی است. تهی شدن خوشه‌ها ازجمله مشکل دیگر این روش است. برخلاف روش سلسله مراتبی این روش تنها خوشه‌بندی را در یک سطح انجام می‌دهد (Zhao, Li, and Cang ۲۰۱۵).
- **مدل‌های مبتنی بر چگالی:** برای پیدا کردن خوشه‌های با شکل مختلف، روش‌های مبتنی بر چگالی ارائه‌شده‌اند. این روش‌ها با این فرض عمل می‌کنند که خوشه‌ها از اشیائی در فضاهای چگالی از داده تشکیل‌شده‌اند که توسط فضاهای خالی از هم جدا می‌شوند. خوشه‌بندی DBSCAN یکی از این نوع روش‌ها است که مبتنی بر روش‌های خوشه‌بندی چگالی بر روی ناحیه‌های متصل شده با استفاده از چگالی بالا عمل می‌کند. مزیت این روش به نسبت روش‌های دیگری خوشه‌بندی مانند خوشه‌بندی کی-میانگین این است که نسبت

[11] Partitioning Method

[2] K-Means

[3] Centroid

به شکل داده‌ها حساس نیست و می‌تواند اشکال غیرمنظم را نیز در داده‌ها تشخیص دهد (Han, Pei, and Kamber ۲۰۱۱). در این روش خوشه‌بندی هر داده متعلق به خوش C در دسترس چگالی سایر داده‌های متعلق به آن خوشه است و در دسترس چگالی هیچ داده دیگری قرار ندارد.

```
for each o ∈ D do
    if o is not yet classified then
        if o is a core-object then
            collect all objects density-reachable from o
            and assign them to a new cluster.
        else
            assign o to NOISE
```

شکل ۲-۷ شبه کد الگوریتم DBSCAN

یکی از تفاوت‌های عمده‌ی این الگوریتم با کا-میانه این است که الگوریتم DBSCAN نیاز به تعیین تعداد خوشه توسط کاربر ندارد و خود الگوریتم می‌تواند گروه‌های یکسان را بر مبانی بر چگالی آن‌ها شناسایی کند. در الگوریتم DBSCAN دو پارامتر اصلی وجود دارد: شعاع (که به آن Epsilon نیز گفته می‌شود) و حداقل نقاط موجود در یک خوشه (که به آن MinPoints نیز گفته می‌شود). نحوه کار الگوریتم به این صورت است که الگوریتم ابتدا یک نمونه (یک نقطه در فضای برداری) را انتخاب می‌کند و با توجه به شعاع به دنبال همسایه برای این نقطه در فضا می‌گردد. درصورتی‌که الگوریتم در شعاع مشخص بتواند به تعداد MinPoint نقطه پیدا کند، در این صورت همه آن نقطه‌ها باهم به یک خوشه تعلق می‌گیرند. الگوریتم سپس به دنبال یکی از نقطه‌های مجاور نقطه فعلی می‌رود تا دوباره با شعاع مشخص‌شده در آن نقطه به دنبال نقاط همسایه دیگر بگردد. درصورتی‌که تعداد نقاط همسایه جدیدی یافت شود، الگوریتم دوباره همه نقاط جدید را با نقاط قبلی در یک خوشه قرار می‌دهد. اگر نقطه‌ی جدیدی در همسایگی پیدا نشد عملیات خوشه‌بندی در این نقاط به اتمام رسیده و خوشه را تشکیل می‌دهد سپس برای پیدا کردن خوشه‌های دیگر در نقاط دیگر، به‌صورت تصادفی یک نقطه دیگر را انتخاب خواهد کرد. شروع به یافتن همسایه و تشکیل خوشه جدید برای آن نقطه می‌کند. این کار تا زمانی که تمام نقاط بررسی شوند و در خوشه‌ای قرار نگیرند ادامه پیدا می‌کند.

- **روش‌های مبتنی بر شبکه**[1]: دسته خاصی از روش‌های مبتنی بر چگالی هستند که در آن‌ها هر منطقه مجزا در فضای داده که جست‌وجو می‌شود، در ساختار شبکه مانندی قرار می‌گیرد. الگوریتم‌های شبکه‌ای اصولاً زمان پردازش سریع‌تری دارند. در ابتدا یک شبکه یکنواخت را برای جمع‌آوری داده‌های منطقه‌ای آماری مهیا می‌شود و سپس به‌جای اینکه به‌طور مستقیم بر روی پایگاه داده کاری انجام دهند، به خوشه‌بندی شبکه می‌پردازند. البته برای توزیع داده‌های بسیار بی‌نظم استفاده از یک شبکه یکنواخت برای به دست آوردن کیفیت دسته‌بندی یا محدودیت زمانی موردنیاز ممکن است کافی نباشد(Liao, Liu, and Choudhary۲۰۰۴).
- **خوشه‌بندی بر مبنای (مبتنی بر) مدل**[2]: این خوشه‌بندی شامل دو بخش مفهومی و شبکه‌های عصبی است. در روش‌های مفهومی، روش‌های مبتنی بر مدل برخلاف روش‌های سنتی خوشه‌بندی که به دنبال اعضای مشابه هستند، یک گام فراتر رفته و علاوه بر این کار ویژگی‌های توصیفی نیز برای هر گروه (خوشه) نیز ارائه می‌دهند که هر گروه نشان‌دهنده یک مفهوم یا کلاس خاصی است. بنابراین خوشه‌بندی مفهومی شامل دو گام اصلی است: خوشه‌بندی؛ مشخصه گذاری هر خوشه. بیشتر خوشه‌بندی‌های مفهومی از مدل‌های آماری (که در آن معیارهای احتمالی استفاده می‌کنند) مورداستفاده قرار می‌دهند. این توضیح‌های احتمالی برای ارائه مفاهیم مورداستفاده قرار می‌گیرند در روش‌های پیشین، فرضیاتی مبنی بر وجود یک توزیع آماری برای داده‌ها وجود نداشت. در روش «خوشه‌بندی بر مبنای مدل» یک توزیع آماری برای داده‌ها فرض می‌شود. هدف در اجرای خوشه‌بندی بر مبنای مدل، برآورد پارامترهای توزیع آماری به همراه متغیر پنهانی است که به‌عنوان برچسب خوشه‌ها در مدل معرفی‌شده است. مدل **ترکیب گوسی**[3]، ازجمله روش‌هایی است که دقیقاً مربوط به آمار است. این روش‌ها بر اساس مدل‌های توزیع عمل می‌کنند. با این فرض عمل می‌کنند

[1] Grid-Based Method

[2] Model-Based Clustering

[3] Gausian Mixture Model

که هر عضو از خوشه‌ها به‌احتمال‌زیاد توزیع یکسانی دارند. خوشه‌ها را بر اساس تابع چگالی احتمال از متغیرهای مشاهده‌شده به‌عنوان ترکیبی از تراکم نرمال چند متغیر تشکیل می‌دهد. خوشه‌ها به صورتی انتخاب می‌شوند که احتمال پسین را بیشینه کنند. این روش به‌عنوان خوشه‌بندی نرم نیز در نظر گرفته می‌شود. احتمال پسین نشان می‌دهد که هر نقطه داده نشان می‌دهد که احتمال وابستگی نیز به خوشه‌های دیگر دارد. احتمال اینکه این روش بهینه محلی قرار گیرد نیز وجود دارد. روش آماری COBWEB و روش‌های **شبکه‌های عصبی** ازجمله روش‌های دیگر این نوع هستند.

- **تحلیل داده‌های خارج از محدوده** [۱]: تشخیص داده‌های پرت می‌تواند به‌عنوان یک مرحله پیش‌پردازش در مسیر داده‌کاوی، و یا مستقلاً به‌عنوان یک عملیات داده‌کاوی مطرح شود. روش‌های متعددی برای تشخیص داده‌های پرت وجود دارد. داده پرت به دادهای اطلاق می‌گردد که معمولاً در یک مجموعه‌داده نسبت به سایر مقادیر موجود بزرگ‌تر یا کوچک‌تر است. داده پرت مشاهده‌ای است که در فاصله دورتری از سایر داده‌ها قرار می‌گیرد و با مقدار مورد انتظاری که داریم متفاوت است.

خوشه‌بندی اسناد کاربردهای متنوعی دارد برخی از این کاربردها که در (Aggarwal and Zhai ۲۰۱۲) به آن اشاره شده است که عبارت‌اند از: سازمان‌دهی اسناد و مرور آن‌ها، خلاصه‌سازی اسناد، خوشه‌بندی اسناد.

۲-۱-۸-۱- خوشه‌بندی «کا-میانه» و «کا-مینه دو بخشی»

روش کا-میانه یکی از روش‌های خوشه‌بندی همه منظوره است که به‌صورت گسترده مورداستفاده قرار می‌گیرد (Aggarwal and Zhai۲۰۱۲). هرکدام از این روش‌ها مزایا و معایب خاص خود را دارا هستند. به‌صورت ساده روش کا-میانه پایه را می‌توان به‌صورت زیر بیان کرد:

۷- تعدادی (k تعداد) مرکز خوشه به‌صورت تصادفی در محدوده عناصر انتخاب کن.

۸- عناصر مجموعه‌داده را به نزدیک‌ترین مرکز خوشه نسبت بده.

---

[۱] Outlier Detection

۹- مراکز خوشه را مجددا محاسبه کن.

۱۰- مرحله ۲ و ۳ را تا زمانی که مراکز خوشه تغییری نکردند ادامه بده.

ازآنجاکه روش کامیانه سرعت بالاتری نسبت

در بسیاری مقالات برای خوشه‌بندی اسناد و متون از دو روش خوشه‌بندی سلسله مراتبی و خوشه‌بندی کا-میانه استفاده‌شده است. مطالعات نشان داده است که باوجوداینکه روش سلسله مراتبی دقت بالاتری دارد ولی روش کا-مینه سرعت بالاتری دارد. تغییرات بسیاری برای بهبود عملکرد روش کا-میانه ارائه‌شده است (برای مطالعه بیشتر به (Jain۲۰۱۰) مراجعه شود). در پژوهش‌های مختلف نشان داده‌شده است که روش «کا-مینه دوبخشی» روشی است که از کا-میانه بهتر عمل می‌کند و دقتی همانند روش سلسله مراتبی یا حتی بهتر از آن دارد. (Steinbach, Karypis, and Kumar ۲۰۰۰ .Wei et al ,۲۰۱۵). روش خوشه‌بندی «کا-مینه دوبخشی» به‌صورت زیر عمل می‌کند:

۱- یک خوشه را انتخاب کن.

۲- با استفاده از الگوریتم کا-میانه پایه خوشه را به دو زیر خوشه تقسیم کن.

۳- برای تعداد مشخصی تکرار گام ۲ را برای رسیدن به بهترین خوشه تکرار کن.

۴- گام‌های ۱،۲ و ۳ را تا زمان رسیدن به تعداد خوشه موردنظر تکرار کن.

روش‌های مختلفی برای انتخاب خوشه‌ای که قرار است به دو قسمت تقسیم بشود وجود دارد. انتخاب خوشه بعدی می‌توان بر اساس بزرگی اندازه خوشه‌های موجود، خوشه‌ای که کمترین میزان شباهت را اعضای آن دارند، یا بر اساس ترکیت پارامترهای میزان شباهت و اندازه باشد. تحقیقات نشان داده است که اختلاف این پارامترها در خوشه‌بندی تأثیر چندانی در آن نمی‌گذارد (Karypis, Kumar, and Steinbach۲۰۰۰).

به خاطر عملکرد بهتر «کا-مینه دوبخشی» نسبت به کا-میانه، علاوه بر خوشه‌بندی متون در کاربردهایی که از مدل «کیسه‌ای از کلمات[۱]» استفاده می‌شود کاربرد دارد، برای مثال در (Zhao, Li, and Cang۲۰۱۵) در سال ۲۰۱۵ با استفاده از روش کا-میانه دوبخشی و روش BOW تشخیص

[۱] Bag of Word (BOW)

چهره انجام‌گرفته است. بر اساس پژوهش انجام‌شده در (Khazaei and Ghasemzadeh ۲۰۱۵) در سال ۲۰۱۵ نشان داده‌شده است که نتایج گرفته‌شده از خوشه‌بندی کا-میانه از پژوهش‌های انجام‌گرفته در متن‌های انگلیسی قابل توسعه برای متن‌های فارسی نیز است. در برخی تحقیقات نشان داده‌شده است که روش کا-میانه با الگوریتم ژنتیک در داده‌های غیر متنی نتیجه بهتری می‌دهد (Banerjee, Choudhary, and Pal ۲۰۱۵). در پژوهش (Steinbach, Karypis, and Kumar ۲۰۰۰) در سال ۲۰۰۰ نشان داده است که خوشه‌بندی کا-میانه دوبخشی بهتر از روش‌های خوشه‌بندی کا-میانه (سرعت‌بالا در میان روش‌های خوشه‌بندی) و سلسله مراتبی (دارای کیفیت بهتر خوشه‌بندی نسبت به کا-میانه عادی) دارد مقایسه و ارزیابی صورت گرفته بر اساس شاخص آنتروپی بوده است.

۲-۱-۸-۲- خوشه‌بندی FCM[۱]

روش خوشه‌بندی که یک تابع هدف را حداقل کند متعلق به دسته الگوریتم‌های تابع هدف هستند. زمانی که تابع هدف الگوریتمی برای حداقل کردن خطا باشد به‌اصطلاح آن را C-Means گفته می‌شود که C نشان‌دهنده تعداد کلاس‌ها یا خوشه‌هاست. درصورتی‌که برای این منظور از هر نوع روش فازی استفاده شود به آن FCM گفته می‌شود (Nayak, Naik, and Behera ۲۰۱۵).

روش FCM یکی از بهترین روش‌های خوشه‌بندی پراستفاده است. این نوع خوشه‌بندی یکی از روش‌های خوشه‌بندی است که به تک‌تک داده‌ها اجازه می‌دهد که عضو یک یا چند تا از خوشه‌ها بشوند. این روش یکی از روش‌های متداول در تشخیص الگو است. خوشه‌بندی فازی را می‌توان یکی از روش‌های خوشه‌بندی نرم در نظر گرفت. بنابراین هر عضو با درجه‌ای از عضویت مشخص، می‌تواند به تمام خوشه‌ها متعلق باشند. خوشه‌بندی فازی شامل مراحل زیر است:

۱. مقداردهی اولیه به متغیرها (مقدار تابع تعلق و مراکز خوشه را به‌صورت تصادفی قرار بده)
۲. بر اساس تابع تعلق مرکز خوشه جدید راه محاسبه کن. برای تمام عضوهای یک خوشه از رابطه زیر استفاده می‌شود.

[۱] Fuzzy C-Means

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \, x_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad (۲۸-۲)$$

۳. مقدار تابع تعلق را به‌روز کن.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - c_j||}{||x_i - c_k||} \right)^{\frac{2}{m-1}}} \qquad (۲۹-۲)$$

۴. درصورتی‌که تغییری در تابع تعلق مشاهده نشد و یا میزان خطا یا تغییر خطا کمتر از مقداری شد عملیات را به اتمام برسان

مقدار u_{ij} میزان تعلق عضو x_i به خوشه c_j است.

در بسیاری از روش‌ها مقدار ۱,۵ تا ۲,۵ را برای پارامتر فازی بودن (m) انتخاب و پیشنهاد می‌کنند (Pal and Bezdek ۱۹۹۵). افزایش این پارامتر باعث تحمل‌پذیری این الگوریتم در مقابل نویز و داده‌های پرت است (Wu ۲۰۱۲).

۲-۱-۸-۳- مدل‌های نمایش اسناد

یکی از ساده‌ترین روش‌های نمایش اسناد برای الگوریتم‌های خوشه‌بندی استفاده از **VSM** است. در این روش هر سند یک بردار با d عنصر (در فضای کلمات) در نظر گرفته شده است. روش‌های مختلفی برای نشان دادن ویژگی‌ها برای خوشه‌بندی وجود دارد که عبارت‌اند از: قدرت کلمه، تکرار تکرار در سندها، امتیازدهی بر اساس آنتروپی و سهیم بودن واژه (Aggarwal and Zhai ۲۰۱۲).

به‌صورت ساده برای نمایش ویژگی‌های بردار از «تکرار کلمات[۱]» استفاده می‌شود. در حالتی دیگر، برای هر بردار از TF-IDF استفاده‌شده است. تحقیقات نشان داده‌شده است که زمانی که تعداد

[۱] Term Frequency (TF)

اسناد بیش از ۱۰۰۰ سند باشند تفاوت چندانی در انتخاب پارامتر بین این دو گزینه ندارد (Larsen and Aone۱۹۹۹).

طول بردار نمایش نیز در خوشه‌بندی تأثیر می‌گذارد در (Larsen and Aone۱۹۹۹) نشان داده است که هرچه طول بردار بیشتر باشد دقت نیز بالاتر می‌رود. اما این امر باعث افزایش زمان پردازش خواهد شد. درصورتی‌که از الگوریتم‌های خوشه‌بندی استفاده می‌کنیم شباهت بین دو سند باید موردمحاسبه قرار گیرد. بیشترین و متداول‌ترین روش برای اندازه‌گیری شباهت استفاده از تابع «کوسین[۱]» است. این تابع به‌صورت زیر محاسبه می‌شود:

$$cosine\left(d_1, d_2\right) = \frac{d_1 . d_2}{\left\|d_1\right\| \left\|d_2\right\|} \qquad (۳۰-۲)$$

معیار کمی خبره‌تر برای پیدا کردن شباهت استفاده از ضریب همبستگی پیرسون است. ضریب همبستگی معیاری است برای این که دو مجموعه‌داده چقدر روی یک خط راست قرار دارند. فرمول محاسبه این ضریب از فاصله اقلیدسی پیچیده‌تر است، اما نتایج بهتری را برای زمانی که داده نرمال شده نیست می‌دهد. برای مثال در خوشه‌بندی متون وقتی سندها دارای تعداد کلمات متفاوتی باشند، ممکن است فاصله اقلیدسی نتیجه درستی به ما ندهد اما ضریب همبستگی نتایج بهتری در این شرایط خواهد داد (Segaran۲۰۰۷).

$$r = \frac{\sum XY - \frac{1}{N} * \sum X * \sum Y}{\sqrt{\left(\sum X^2 - \frac{1}{N}(\sum X)^2\right) * \left(\sum Y^2 - \frac{1}{N} * (\sum Y)^2\right)}} \qquad (۳۱-۲)$$

انتخاب خوشه‌ها یکی از بزرگ‌ترین مسئله‌هایی است که در خوشه‌بندی با آن مواجه هستیم. معمولاً برای حل این مشکل خوشه‌بندی با مقدار مختلفی انجام می‌شود و تعداد خوشه‌ای که بهترین

[۱] Cosine

نتیجه را داده است به‌عنوان تعداد خوشه انتخاب می‌شود. بهترین نتیجه خوشه‌بندی نیز به از شاخص‌های مختلفی قابل‌محاسبه است که در بخش بعدی به آن خواهیم پرداخت. روش دیگر نمایش کلمات و اسنادی استفاده از روش‌های **بازنمایی کلمه** است که در بخش بعدی به‌تفصیل شرح داده‌شده است.

۲-۱-۸-۴- بررسی ادبیات خوشه‌بندی اسناد

ازآنجاکه کلماتی که در فضای برداری (بازنمایی کلمات) در کنار هم هستند ازنظر معنایی کلمات به هم مربوط هستند (Mikolov et al. ۲۰۱۳)، خوشه‌بندی این کلمات می‌تواند برای کشف گروه‌های هم‌معنی مورداستفاده قرار گیرد (Wang et al. ۲۰۱۵).

در سال ۱۹۹۹ روشی با ترکیب دو روش خوشه‌بندی برای خوشه‌بندی متن ارائه کرده است. در روش ارائه‌شده بعد از دو مرحله انتخاب ویژگی سه روش برای انتخاب مراکز خوشه (تصادفی، باکشات[۱]، تقسیم کردن[۲]) در هر مرحله از خوشه‌بندی پیشنهادشده است . این پژوهش نشان داده است که در روش‌های خوشه‌بندی انتخاب و به‌روزرسانی مرکز مؤثرتر از روش انتخاب تصادفی مراکز در ابتدای کار است (Larsen and Aone ۱۹۹۹).

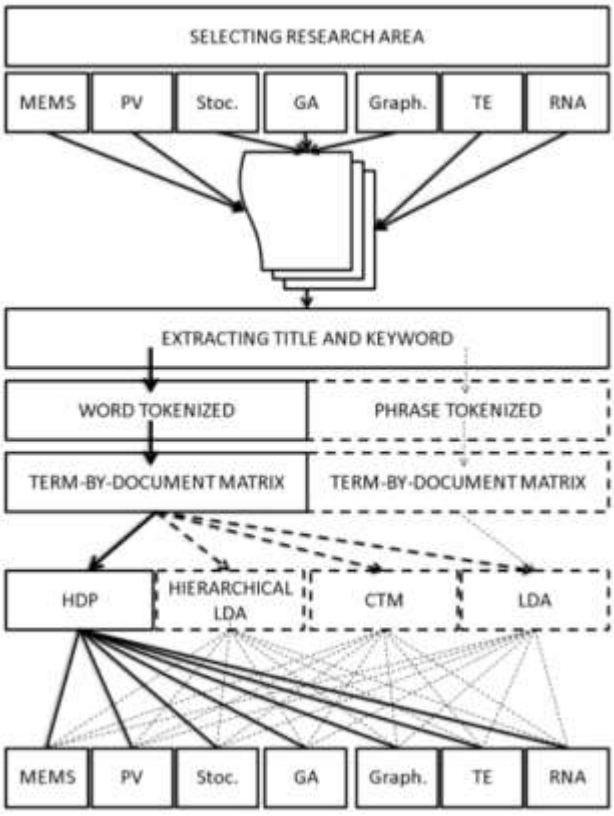
در ۲۰۱۱، «سماه فوداه» در (Fodeh, Punch, and Tan ۲۰۱۱) با استفاده از آنتولوژی برای رفع ابهام کلمات (WSD) اقدام کرده و استخراج روابط معنایی آن‌ها را در خوشه‌بندی لحاظ کرده است. بعد از اینکه معانی و مجموعه‌های بالاسری (ریشه) کلمات استخراج شدند برای کاهش بیشتر کلمات (ویژگی‌ها) از بهره اطلاعاتی[۳] استفاده می‌کند. در حقیقت ویژگی‌هایی مناسب هستند که بعد از جایگزین شدن با مفهومشان هنوز IG بالایی داشته باشند. درصورتی‌که با جابجایی کلمه با مفهوم، کاهش میانگین آنتروپی مشاهده شود، نشان‌دهنده اینست که بهره اطلاعاتی کلمه از مفهوم آن بالاتر است. در خوشه‌بندی ابتدایی کلمات

[۱] Buckshot

[۲] Fractionation

[۳] Information Gain (IG)

«چی یائو» در سال ۲۰۱۴ بر اساس مدل‌سازی موضوعی[۱] اسناد علمی استخراج‌شده از WoS را خوشه‌بندی کرده است (۱۲۵۴ مقاله). در این پژوهش از کلماتی تکی و عبارات چندکلمه‌ای به‌عنوان کلمات کلیدی استفاده کرده است که نتایج نشان داده است که کلمات تکی کاهش عملکردی نسبت به عبارات چندکلمه‌ای نداشته‌اند. بعد از اجرای الگوریتم‌های مختلف چند کلمه برتر هر موضوع را به‌عنوان نماینده آن موضوع انتخاب کرده است. موضوعی که بیشترین درصد و بالای ۵۰ درصد از یک سند را داشته باشد را به‌عنوان موضوع آن مقاله در نظر می‌گیرد. مقاله را نیز در دسته‌ای که موضوعش است قرار می‌دهد. لازم به ذکر است که برای موضوع‌های تشکیل‌شده هم به‌صورت دستی اسم‌گذاشته است (Yau et al. ۲۰۱۴).



شکل ۲-۸ خوشه‌بندی با استفاده از روش‌های مدل‌سازی موضوعی (Yau et al. ۲۰۱٤)

[۱] Topic Modeling

«تینگ تینگ وی» روش خوشه‌بندی در سال ۲۰۱۴ بر اساس روابط معنایی بین کلمات ارائه کرده است. این روش با استفاده از WordNet و آنتولوژی خوشه‌بندی را انجام می‌دهد. برای این منظور یک زنجیره‌ای از لغات[۱] برای هر سند درست می‌کند که محتوای متن را نشان می‌دهد. با این روش روابط معنایی مانند متضادها، مترادف‌ها و دیگر روابط در خوشه‌بندی دخیل می‌شوند. با استفاده از زنجیره کلماتی که برای هر سند تشکیل می‌شود، برچسب توضیحی برای آن سند نیز زده می‌شود. ازآنجاکه دخیل کردن روابط معنایی مانند هم‌معنی‌ها و چند معنی‌ها که برای رفع ابهام در متن مورداستفاده می‌شود باعث بروز تنک شدن مجموعه‌داده می‌شود اما این روش ادعا کرده است که با تشکیل زنجیره لغات این عیب را برطرف کرده است (Wei et al. ۲۰۱۵).

همبستگی[۲] کلمات عبارت‌اند از روابط بین کلمات که بخش‌های مختلف متن را نشان می‌دهند.

زنجیره لغات[۳] توالی از کلمات به هم مرتبط هستند که درک مهمی در خصوص مفهوم محتوای متن ارائه می‌دهند. بنابراین محاسبه زنجیره لغات این امکان را می‌دهد که موضوع اصلی متن را تشخیص دهیم.

با بررسی الگوریتم‌های خوشه‌بندی مختلف و ارزیابی آن‌ها در پژوهش (Fahad et al. ۲۰۱۴) که در سال ۲۰۱۴ انجام شد، نشان داده شده که الگوریتم‌های خوشه‌بندی با تکرارهای مختلف به خاطر برخی خواص تصادفی بودنشان، نتایج یکسانی نمی‌دهند و در حالت پایدار[۴]ی قرار ندارند. پیشنهادشده است که از حالت ترکیبی از خوشه‌بندی‌ها برای رفع این مشکل استفاده شود. شکل زیر برخی از الگوریتم‌های خوشه‌بندی را به‌صورت مختصر با یکدیگر مقایسه کرده است (Fahad et al. ۲۰۱۴).

---

[۱] Lexical Chain

[۲] Cohesion

[۳] Lexical Chain

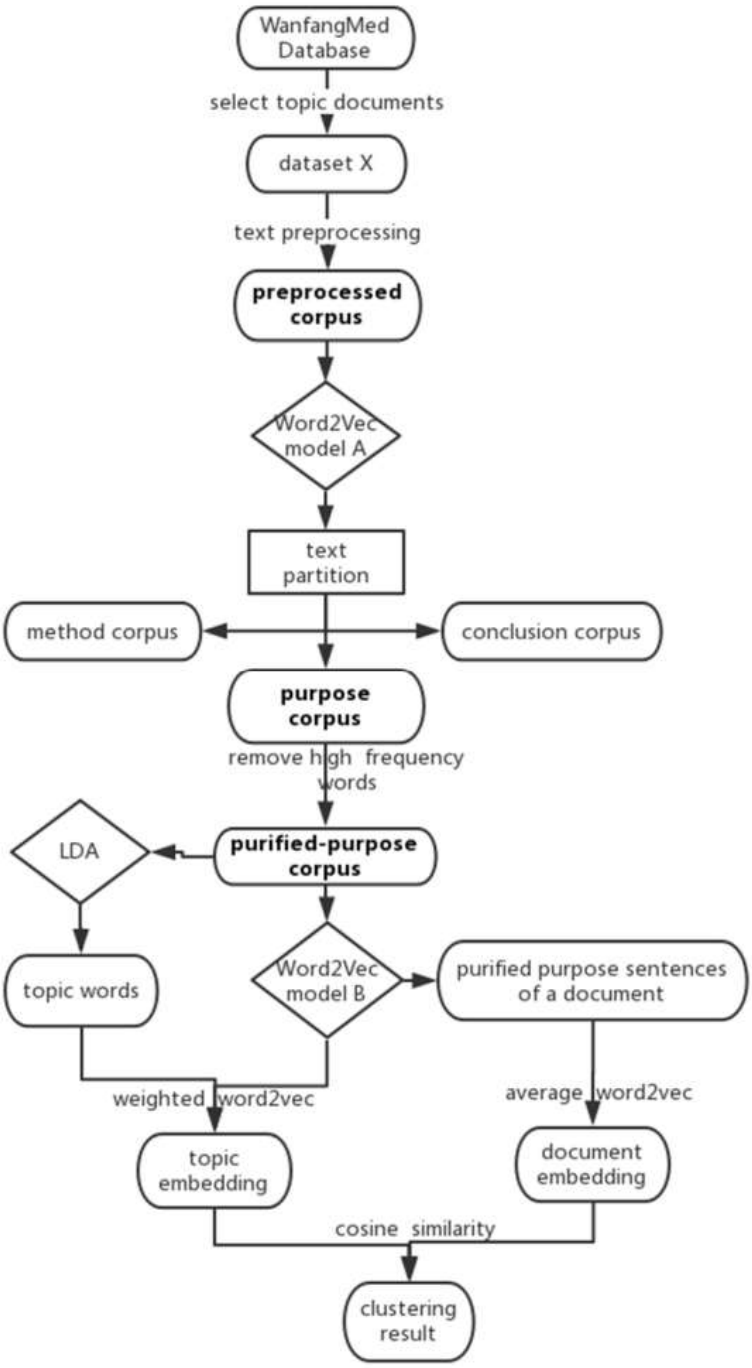
[۴] Stability

روش برای استخراج کلیدواژه‌ها و خوشه‌بندی در سال ۲۰۱۵ در (Habibi and Popescu-Belis ۲۰۱۵) ارائه شد. در این پژوهش در مرحله اول کلیدواژه‌ها را از متن استخراج کرده (که بیشترین پوشش را از موضوع‌های شناسایی‌شده در متن دارند) و در مرحله بعدی کلیدواژه‌های استخراج‌شده را خلاصه می‌کند. برای استخراج کلیدواژه‌ها از روش‌های مدل‌سازی موضوعی استفاده کرده است و چند کلمه اول هر موضوع را به‌عنوان کلیدواژه استخراج کرده است تا دامنه بیشتری از کلیدواژه‌ها انتخاب شوند و در بازیابی و خوشه‌بندی تنوع بیشتری داشته باشد.

سال ۲۰۱۵ «خزاعی» در (Khazaei and Ghasemzadeh۲۰۱۵) به مقایسه عملکرد الگوریتم کا-میانه در متن‌های فارسی و انگلیسی پرداخته است. این عملکرد تنها در پارامتر میانگین مجموع مربع خطاها صورت پذیرفته است ازاین‌رو نمی‌توان گفت که نتایج به‌صورت کامل و جامعی هستند. این پژوهش نشان داد که عملکرد الگوریتم کا-میانه در دو زبان مشابه هم می‌تواند باشد. بعد از انتخاب ویژگی با روش PCA به استخراج ویژگی انجام داده است و مشاهده شد که به خاطر تفاوت در ساختار زبان فارسی نسبت به زبان انگلیسی ویژگی‌های بیشتری از زبان فارسی استخراج‌شده است.

سال ۲۰۱۸ «چان ژونگ لی» روشی با استفاده از ترکیب vec۲word و LDA برای خوشه‌بندی چکیده‌های علمی ارائه کرده است. این روش به‌منظور رفع مشکل تنک بودن کلیدواژه‌ها (در پی آن غیر پایداری و غیردقیق بودن خوشه‌بندی) و به خاطر کوتاه بودن متن دامنه‌های کوچک علمی به خاطر همپوشانی زیاد کلمات غیر مهم باهم غیر قابل‌کشف و یا حتی باعث حذف کلمات می‌شوند. در گام اول کلیدواژه‌های مهم را از بخش ابتدایی چکیده *به دست* آورده می‌شود و در ضمن کلمات با تکرار بالا برای استخراج کلیدواژه‌های فوق حذف خواهند شد. مدل‌های LDA و Vec۲Word آموزش داده و باهم ترکیب می‌شوند. درنهایت با استفاده از این دو مدل عملیات خوشه‌بندی انجام خواهد شد. (استفاده از چکیده بجای کلیدواژه باعث رفع ابهام کلمات و بهبود در خوشه‌بندی خواهد شد)(Li et al. ۲۰۱۸) بنابراین این روش برای کاهش تأثیر اندازه کوچک متن از ترکیب مدل‌های LDA و Vec۲Word استفاده کرده است (این ادعا را دارد که ترکیب این دو باعث کاهش تنک بودن متن خواهد شد) و برای دامنه‌های کوچک تحقیقاتی از جملات هدفمند در متن چکیده (بخش اول از سه بخش چکیده) استفاده کرده است. در ضمن برای هر بار پردازش متن‌ها یک مجموعه

کلمات‌ایست جدید درست می‌کند. تفاوتی که این پژوهش با پژوهش (Wang, Ma, and Zhang ۲۰۱۶) دارد استفاده از چکیده خالص شده و جایگزینی فاصله کوسین با اقلیدسی است (Li et al. ۲۰۱۸).



شکل ۲-۹ روش ارائه‌شده در خوشه‌بندی ارائه‌شده در مرجع (.Li et al ۲۰۱۸)

۲-۱-۸-۵- روش‌های ارزیابی خوشه‌ها

مسئله‌ی ارزیابی الگوریتم‌ها به‌صورت کلی یکی از مهم‌ترین و اصلی‌ترین زمینه‌های هوش‌مصنوعی، تحقیقات علمی، یادگیری ماشین، کشف دانش و داده‌کاوی است (Kou, Peng, and Wang ۲۰۱۴). روش‌های نظارت شده می‌توانند براساس معیارهای دقت[۱] و صحت[۲] ارزیابی شوند، اما روش‌های خوشه‌بندی به خاطر ذات تحلیل خوشه دشوارتر هستند.

اعتبارسنجی خوشه یک شیوه‌ی (روال فرمالی) است که برای ارزیابی نتایج *به‌دست‌آمده* از خوشه‌بندی استفاده می‌شود. این شیوه‌فنی برای پیدا کردن مجموعه‌ای از بهترین خوشه‌ها که به‌صورت طبیعی و بدون اطلاعات کلاس، داده‌ها را تقسیم می‌کنند مورداستفاده قرار می‌گیرد. نتایج حاصل از اعمال الگوریتم‌های خوشه‌بندی روی یک مجموعه‌داده با توجه به پارامترهای مختلف الگوریتم انتخاب‌شده، می‌تواند بسیار متفاوت از یکدیگر باشد. این روال هم به‌صورت کمی و بی‌طرف می‌بایست انجام شود.

شاخص‌های ارزیابی و اعتباربخشی خوشه‌بندی بر اساس سه معیار مختلف تعریف می‌شوند: درونی، نسبی، و خارجی (Jain and Dubes ۱۹۸۸). هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب را با داده‌های موردنظر داشته باشند. دو معیار اندازه‌گیری برای ارزیابی و انتخاب یک طرح خوشه‌بندی بهینه ارائه‌شده است که معمولاً شاخص‌های ارزیابی از یکی یا ترکیبی از هر دو معیار استفاده می‌کنند(Rendón et al. ۲۰۱۱):

۱- فشردگی[۳]: اعضای هر خوشه باید تا آنجا که می‌شود به یکدیگر نزدیک باشند. یک روش اندازه‌گیری متداول برای این معیار واریانس است.

۲- تفکیک‌پذیری[۴]: خوشه‌ها خود بایستی به‌اندازه کافی از یکدیگر جدا باشند. سه روش برای اندازه‌گیری میزان جدایی خوشه‌ها که مورداستفاده قرار می‌گیرد که عبارت‌اند از: فاصله

---

[۱] Precision

[۲] Accuracy

[۳] Compactness

[۴] Separation

بین نزدیک‌ترین عضوهای خوشه‌ها، فاصله بین دورترین عضو خوشه‌ها و فاصله بین مراکز خوشه‌ها.

همانطر که گفته شد سه روش مختلف برای ارزیابی نتایج الگوریتم‌های خوشه‌بندی وجود دارد (Kovács, Legány, and Babos ۲۰۰۵):

۱- معیار خارجی

۲- معیار داخلی

۳- معیار نسبی

هر دو معیار داخلی و خارجی بر مبنای روش‌های آماری هستند. این روش‌ها توان محاسبانی بالایی نیاز دارند. ارزیابی خارجی بر اساس دانش قبلی که از داده‌ها وجود دارد انجام می‌شود. ارزیابی داخلی نیز بر مبنای اطلاعات ذاتی موجود در داده‌ها انجام خواهد شد. مشکل این دو روش پیچیدگی محاسباتی موردنیاز آن‌ها است. در معیار نسبی نیز بر اساس مقایسه خروجی‌های دو الگوریتم خوشه‌بندی، الگوریتم بهتر مورد انتخاب قرار می‌گیرد. معیارهایی که در ادامه به آن‌ها پرداخته می‌شود برای الگوریتم‌های خوشه‌بندی محکم[۱] مورداستفاده قرا می‌گیرند. البته لازم به ذکر است که برخی از روش‌های خوشه‌بندی به خاطر نداشتن مرکز خوشه قابلیت ارزیابی داخلی را به‌صورت ذاتی ندارند (Fahad et al. ۲۰۱۴).

۲-۱-۸-۵-۱- شاخص‌های خارجی

در ادامه به بررسی برخی از شاخص‌های خارجی در خوشه‌بندی خواهیم پرداخت.

۲-۱-۸-۵-۱-۱- معیار F

این شاخص از ترکیب مفهوم بازیابی[۲] و دقت[۳] که در بازیابی اطلاعات استفاده می‌شوند *به‌دست‌آمده* است. مقدار Recall و دقت از روابط زیر *به دست* می‌آیند:

---

[۱] Crisp

[۲] Recall

[۳] Precision

$$Recall(i,j) = \frac{n_{ij}}{n_i} \qquad (۳۲-۲)$$

$$Precision(i,j) = \frac{n_{ij}}{n_j} \qquad (۳۳-۲)$$

در روابط فوق n_ij تعداد نمونه‌ها از کلاس I هستند که در خوشه j قرار گرفته‌اند و n_j تعداد اعضایی که در خوشه j هستند. معیار F از خوشه j و کلاس i با عبارت زیر محاسبه می‌شود:

$$F(j,j) = \frac{۲Recall(i,j)Preciion(i,j)}{Precision(i,j) + +Recall(i,j)} \qquad (۳۴-۲)$$

مقدار معیار F بین بازه صفر و یک قرار دارد و هرچه عدد آن بزرگ‌تر باشد، کیفیت خوشه‌بندی آن بهتر خواهد بود.

۲-۱-۸-۵-۱-۲- آنتروپی[۱]

از آنتروپی به‌عنوان یک معیاری برای ارزیابی کیفیت خوشه‌ها مورداستفاده قرار می‌گیرد. آنتروپی خالص بودن برچسب‌های کلاس خوشه‌ها را اندازه‌گیری می‌کند. بنابراین اگر هر خوشه تنها شامل یک برچسب کلاس باشد مقدار آنتروپی آن خوشه برابر صفر خواهد بود. از طرفی واضح است که اگر در هر خوشه برچسب‌های متنوعی وجود با فراوانی زیاد داشته باشد مقدار آنتروپی آن افزایش پیدا می‌کند. آنتروپی هر خوشه به‌صورت زیر محاسبه خواهد شد:

$$E_j = -\sum_{i=۱}^{L} p_{ij} \log(P_{ij}) \qquad (۳۵-۲)$$

که در رابطه L تعداد کلاس‌ها

P_{ij}: احتمال اینکه یک عنصر از خوشه j عضو کلاس i باشد.

آنتروپی کل از خوشه‌ها مجموع وزن‌دار از تمام خوشه‌ها است و به‌صورت زیر تعریف می‌شود:

[۱] Entropy

$$E_j = -\sum_{j=1}^{k} \frac{\beta_j}{n} e_j \qquad (۳۶-۲)$$

K: تعداد خوشه‌ها

L: تعداد اسناد موجود در خوشه‌ها است.

در حالت کلی خوشه‌بندی بهتر مقدار آنتروپی کمتری دارد.

۲-۱-۸-۵-۱-۳- خلوص[۱]

معیار خلوص یک خوشه نشان‌دهنده درصد درستی اسنادی است که به‌درستی در خوشه‌های خود قرار گرفته‌اند. این معیار بسیار مشابه مقدار آنتروپی است. مقدار خلوص برابر درصد اسنادی است که به درستی در یک خوشه قرار دارند. مقدار خلوص به‌صورت زیر محاسبه خواهد شد:

$$P_j = \frac{1}{n_j} Max_i(n_i^j) \qquad (۳۷-۲)$$

$$Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j \qquad (۳۸-۲)$$

تعداد اشیائی در خوشه j با برچسب کلاس I.

P_j نشان دهنده این است که نسبت بیشترین تعداد اعضایی که در یک خوشه قرار دارند و عضو یک کلاس هستند.

N تعداد کل نمونه‌ها

n_j تعداد اعضای داخل خوشه jام

M تعداد خوشه‌ها

هرچقدر میزان خلوص بیشتر باشد خوشه‌بندی بهتر انجام‌شده است.

[۱] Purity

۲-۱-۸-۵-۱-۴- معیار ارزیابی آکائیکه[۱]

معیار اطلاعاتی آکائیکه معیاری برای سنجش نیکویی برازش است. این معیار بر اساس مفهوم آنتروپی بناشده است و نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. به‌عبارت‌دیگر، این معیار تعادلی میان دقت مدل و پیچیدگی آن برقرار می‌کند. این معیار توسط «هیروتسوگو آکائیکه» برای انتخاب بهترین مدل آماری پیشنهاد شد.

با توجه به داده‌ها، چند مدل رقیب ممکن است با توجه به مقدار AIC رتبه‌بندی شوند و مدل دارای کمترین AIC بهترین است. از مقدار AIC می‌توان استنباط نمود که به‌عنوان‌مثال سه مدل بهتر وضعیت نسبتاً یکسانی دارند و بقیه مدل‌ها به‌مراتب بدتر هستند، اما معیاری برای انتخاب مقدار آستانه‌ای برای AIC که بتوان مدلی را به‌واسطه داشتن AIC بزرگ‌تر از این مقدار رد کرد وجود ندارد. در حالت کلی ، AIC برابر است با:

$$AIC = ۲ * k - ۲ * \ln(L) \qquad (۳۸-۲)$$

که k تعداد پارامترهای مدل آماری است و L مقدار حداکثر تابع درستنمایی برای مدل برآورد شده است.

با تصحیح مرتبه دو AIC برای نمونه‌های با تعداد کمتر AICc به دست می‌آید:

$$AICc = AIC + ۲ * \frac{k\left(k + ۱\right)}{n - k + ۱} \qquad (۳۹-۲)$$

ازآنجایی‌که برای نمونه‌های با تعداد زیاد AICc به AIC همگرا می‌شود، همیشه باید AICc را به جای AIC بکار برد.

۲-۱-۸-۵-۱-۵- شاخص جاکارد[۲]

[۱] Akaike information criterion

[۲] Jaccard Index

شاخص جاکارد برای اندازه‌گیری شباهت بین دو مجموعه‌داده استفاده می‌شود .شاخص جاکارد مقداری بین ۰ و ۱ دارد. شاخص ۱ بدین معنی است که دو مجموعه‌داده یکسان هستند و شاخص ۰ نشان می‌دهد که مجموعه‌داده‌ها هیچ عنصر مشترکی ندارند. شاخص جاکارد توسط فرمول زیر تعریف می‌شود:

$$J(A,B) = \frac{TP}{TP + FP + FN} \qquad (۴۰-۲)$$

بنابراین هرچه شاخص جاکارد کمتر باشد بهتر است.

۲-۱-۸-۵-۱-۶- شاخص Dic

اندازه‌گیری متقارنDice دو برابر وزن TP است درحالی‌که TN نادیده گرفته می‌شود و برابر با ۱F است - اندازه‌گیری F با $\beta = ۱$:

$$J(A,B) = \frac{۲ * TP}{۲ * TP + FP + FN} \qquad (۴۱-۲)$$

۲-۱-۸-۵-۱-۷- Fowlkes–Mallows

شاخص Fowlkes-Mallows شباهت میان خوشه‌های بازگشتی توسط الگوریتم خوشه‌بندی و معیارهای طبقه‌بندی را محاسبه می‌کند. هر چه مقدار شاخص Fowlkes-Mallows بیشتر باشد خوشه‌ها و معیارهای طبقه‌بندی مشابه هستند. این شاخص را می‌توان با استفاده از فرمول زیر محاسبه کرد:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \qquad (۴۲-۲)$$

۲-۱-۸-۵-۲- شاخص‌های داخلی

در ادامه به بررسی برخی از شاخص‌های خارجی در خوشه‌بندی خواهیم پرداخت.

۲-۱-۸-۵-۲-۱- شاخص کالینسکی-هاراباس[۱]

شاخص کالینسکی-هاراباس به نام معیار نسبت واریانس نیز شناخته می‌شود و به‌صورت زیر محاسبه می‌شود:

$$\mathrm{VRC_k} = \frac{SS_b}{SS_w} * \frac{N-k}{K-۱} \qquad (۴۳-۲)$$

در این رابطه N تعداد داده‌ها، K تعداد خوشه‌ها، SS_b واریانس بین خوشه‌ها، SS_w واریانس داخل خوشه‌ها است. SS_b از رابطه زیر محاسبه می‌شود که در این رابطه

۲-۱-۸-۵-۲-۲- شاخص اطلاعاتی بیزی[۲]

یک ابزار در محاسبات آماری برای انتخاب بهترین مدل در میان تعداد محدودی از مدل‌ها استفاده می‌شود. مدلی که کمترین BIC را دارد به‌عنوان مدل بهتر انتخاب می‌شود. این شاخص به‌صورت زیر محاسبه می‌شود:

$$BIC = -\ln(L) + v * \ln(n) \qquad (۴۴-۲)$$

در رابطه فوق n: تعداد اشیاء، L احتمال پارامترها برای تولید در مدل داده‌ها، V پارامترهای آزادی در مدل گوسین است. این شاخص پیچیدگی مدل و برازنده[۳] بودن داده‌ها را در خود لحاظ کرده است. هرچقدر این شاخص کمتر باشد خوشه‌بندی بهتری انجام‌شده است.

۲-۱-۸-۵-۲-۳- معیار اطلاعاتی آکائیکه[۴]

یک ابزار در محاسبات آماری برای انتخاب بهترین مدل در میان تعداد محدودی از مدل‌ها استفاده می‌شود.

۲-۱-۸-۵-۲-۴- شاخص DB[۵]

---

[۱] Calinski-Harabasz

[۲] Bayesian information criterion (BIC)

[۳] Fit

[۴] Akiake InformationCriterion (AIC)

[۵] Davies-Bouldin (DB)

هدف از این شاخص تشخیص مجموعه‌ای از خوشه‌هاست که به‌صورت مناسبی فشرده و از هم جدا هستند. بر اساس روابط زیر هرچقدر این شاخص کمتر باشد خوشه‌بندی بهتر انجام‌شده است (هرچقدر کوچک‌تر باشد بیشتر از هم جداشده‌اند). شاخص DB به‌صورت زیر محاسبه می‌شود:

$$R_{ij} = \frac{S_i + S_j}{M_{in}} \qquad (۴۵-۲)$$

M_ij مقدار جدایی دو خوشه I,j است که هرچقدر بیشتر باشد بهتر است.

S_i مقدار توزیع و پراکندگی درون خوشه I است که هرچقدر کمتر باشد بهتر است.

$$D_i = \max(R_{ij}) \ , i \,!= j \qquad (۴۶-۲)$$

$$DB = \frac{۱}{N}\sum_{i=۱}^{N} D_i \qquad (۴۷-۲)$$

در این رابطه مقدار D_i در بدترین حالت محاسبه می‌شود، هرچقدر مقدار این شاخص کمتر باشد خوشه‌بندی بهتر انجام‌شده است. روش دیگر محاسبه این شاخص به‌صورت زیر است

$$DB = \frac{۱}{N}\sum_{i=۱}^{N} \frac{Max_{ij}\left(d(X_i) + D(X_j)\right)}{d(C_i, C_j)} \qquad (۴۸-۲)$$

۲-۱-۸-۵-۲-۵- شاخص «نیم‌رخ» یا «سیلوهت[۱]»

یکی دیگر از روش‌های ارزیابی خوشه‌بندی، معیار سیلوهت است. این معیار هم به پیوستگی[۲] درون خوشه‌ها و هم به میزان تفکیک‌پذیری آن‌ها بستگی دارد. مقدار نیم‌رخ برای هر نقطه، میزان تعلق آن را به خوشه‌اش در مقایسه با خوشه مجاور اندازه می‌گیرد.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i) \text{ و } b(i))} \qquad (۴۹-۲)$$

---

[۱] Silhouette

[۲] Cohesion

B(i): میانگین فاصله نمونه I با تمام نمونه‌هایی که در خوشه‌های دیگر قرار دارند.

A(i): میانگین فاصله بین نمونه I با تمام نمونه‌هایی که در همان خوشه قرار دارند.

شاخص سیلوهت بین منفی یک و یک قرار دارد و هرچه این شاخص به ۱ نزدیک‌تر باشد کیفیت خوشه‌بندی بهتر خواهد بود. اگر شاخص برای یک گره برابر صفر شد یعنی آن گره می‌تواند بدون اینکه خدشه‌ای به کارایی کل وارد کند، عضو خوشه دیگری شود. اگر مقدار این شاخص کمتر از یک شد به این معنیست که خوشه‌بندی مناسبی برای آن گره انجام‌نشده و آن گره مناسب آن خوشه نیست (de Amorim and Hennig ۲۰۱۵). طبق تحقیقات انجام‌شده این شاخص از بسیاری از شاخص‌های دیگر بهتر عمل می‌کند (de Amorim and Hennig ۲۰۱۵).

۲-۱-۸-۵-۲-۶- شاخص Dunn

«خانواده شاخص‌های دان» (Dunn family indices) در سال ۱۹۷۴ طی مقاله‌ای با توجه به مفهوم «فشردگی[۱]» و «تفکیک‌پذیری[۲]» توسط دان معرفی شدند. او با دو معیار «فاصله» و قطر[۳]، میزان فشردگی و تفکیک‌پذیری را محاسبه کرد.

شاخص دان بر اساس نسبت کمترین فاصله بین خوشه‌ها (که به پراکندگی خوشه‌ها را تخمین میزند)، و بیشترین قطر خوشه (که همبستگی خوشه را تخمین می‌زند) محاسبه می‌شود. این امر می‌تواند باعث شود که داده‌ها نویزی در محاسبه این پارامتر تأثیر بگذارند (de Amorim and Hennig ۲۰۱۵).

$$D(C_i, C_j) = \min_{\in C_i} d(x,y) \;\; x \in \mathrm{C_j}, y \qquad (۲-۵۰)$$

$$diam(C_i) = \mathrm{Max}\, d(x,y) \;\; x,y \in C_i \qquad (۲-۵۱)$$

[۱] Compactness

[۲] Separation

[۳] Diameter

$$Dunn = \frac{\min D(C_i, C_j)}{\max diam(C_i)} \qquad (۵۲-۲)$$

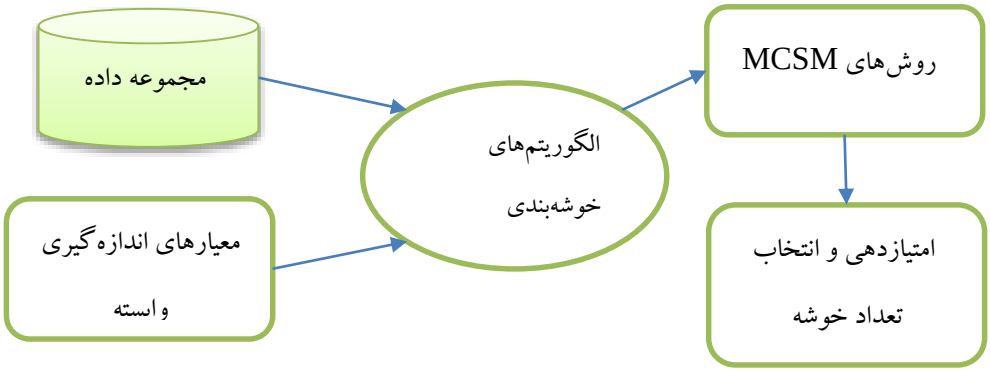
مقدار بیشتر این شاخص نشان‌دهنده کیفیت بهتر خوشه‌بندی است. هرچقدر از یک بزرگ‌تر باشد نشان دهنده این است که خوشه‌ها فشرده‌تر و فاصله بین خوشه‌ها بیشتر است. عیب این روش این است که وقتی تعداد کلاسترها و ابعاد داده بیشتر می‌شوند هزینه محاسبه نیز بسیار بیشتر خواهد شد.

پژوهشی در سال ۲۰۱۱ در (Rendón et al. ۲۰۱۱) نشان‌داده است که معیارهای ارزیابی خوشه درونی در مقایسه با معیارهای ارزیابی خوشه بیرونی در تعیین یک ساختار خوشه‌بندی بهتر عمل می‌کنند. و پژوهش (Liu et al. ۲۰۱۰) در سال ۲۰۱۰ نشان داده است که روش‌های سیلوهت، دان، دی‌بی در حالت‌هایی که توزیع داده‌ها یکنواخت، نویزدار، با چگالی‌های متفاوت و با خوشه‌هایی با اندازه‌های متفاوت دارند بهتر از روش‌های دیگر عمل می‌کنند.

۲-۱-۸-۶- پیدا کردن تعداد بهینه خوشه‌ها برای سندها

مسئله ارزیابی خوشه‌ها یکی از مهم‌ترین و ضروری‌ترین مسئله‌های خوشه‌بندی است. ارزیابی خوشه‌ها بدین منظور صورت می‌گیرد که نشان دهیم خوشه‌ها به‌صورت تصادفی ایجاد نشده‌اند (Jain, Murty, and Flynn ۱۹۹۹). یکی از اساسی‌ترین گام‌های تحلیل خوشه‌بندی، ارزیابی خوشه‌بندی است. دو موضوع اصلی که در ارزیابی خوشه‌بندی باید به آن توجه کرد: یکی تخمین تعداد خوشه‌ها برای مجموعه‌داده است، و دوم ارزیابی الگوریتم‌های خوشه‌بندی است.

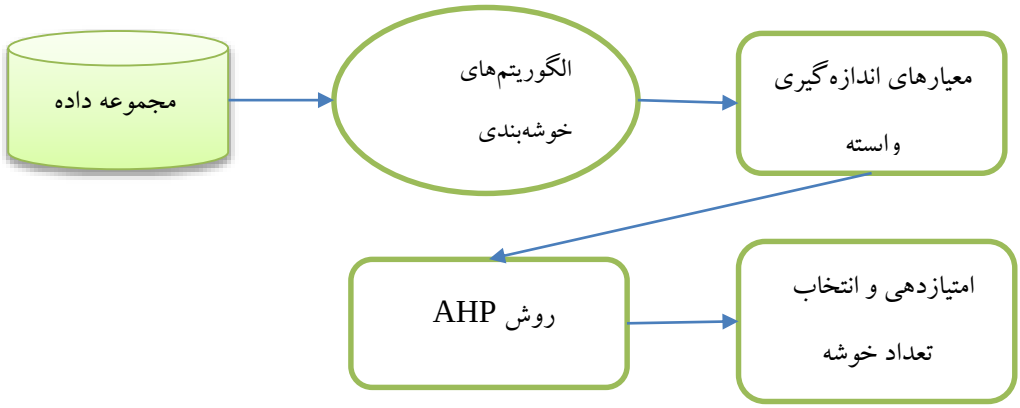
در سال ۲۰۱۲ «پنگ» (Peng, Zhang, Kou, and Shi ۲۰۱۲) روشی ارائه داده است که با استفاده از روش‌های تصمیم‌گیری چند مقداره از میان چندین شاخص ارزیابی خوشه‌بندی تعداد خوشه‌ها را برای یک مجموعه‌داده تخمین می‌زند. در این روش تعداد خوشه‌ها به‌عنوان گزینه‌های مورد تصمیم در نظر گرفته می‌شوند و معیارهای هر یک از خوشه‌ها به‌عنوان معیارهای ارزیابی سیستم تصمیم‌گیری چند مقداره در نظر گرفته می‌شوند.



شکل ۲-۱۰ انتخاب تعداد خوشه‌ها به روش چند مقداره

بنابراین در این روش تعداد خوشه‌های مختلف بر اساس ارزش معیارهای ارزیابی خوشه یک ارزیابی کلی خواهند شد و بالاترین امتیاز به‌عنوان تعداد خوشه‌ی برتر انتخاب می‌شود. این روش از ترکیب سه روش MCDM برای تصمیم‌گیری بهتر استفاده کرده است. برای خوشه‌بندی از روش کا-میانه استفاده‌شده است. یکی از مشکلات این روش پیدا کردن وزن مناسب برای وزن دهی به ویژگی‌ها است. مشکل دیگر این روش استفاده از روش‌های MCDM می‌توان جواب‌های مختلفی را به همراه بیاورد. ایراد دیگر این روش استفاده از خوشه‌بندی قطعی و خشک [۱] است.

روش دیگر ارائه‌شده توسط «آکگول» که در (Akogul and Erisoglu ۲۰۱۷) ارائه‌شده است با استفاده از روش خوشه‌بندی مبتنی بر مدل و مدل تحلیل فرایند سلسله مراتبی (AHP) و با بررسی ۵ شاخص بهترین تعداد خوشه را محاسبه کرده است. شکل زیر نشان‌دهنده کلیات این روش است.



- مرحله اول: ساختار سلسله‌مراتبی AHP ساخته می‌شود. در این مرحله هدف برابر تعداد

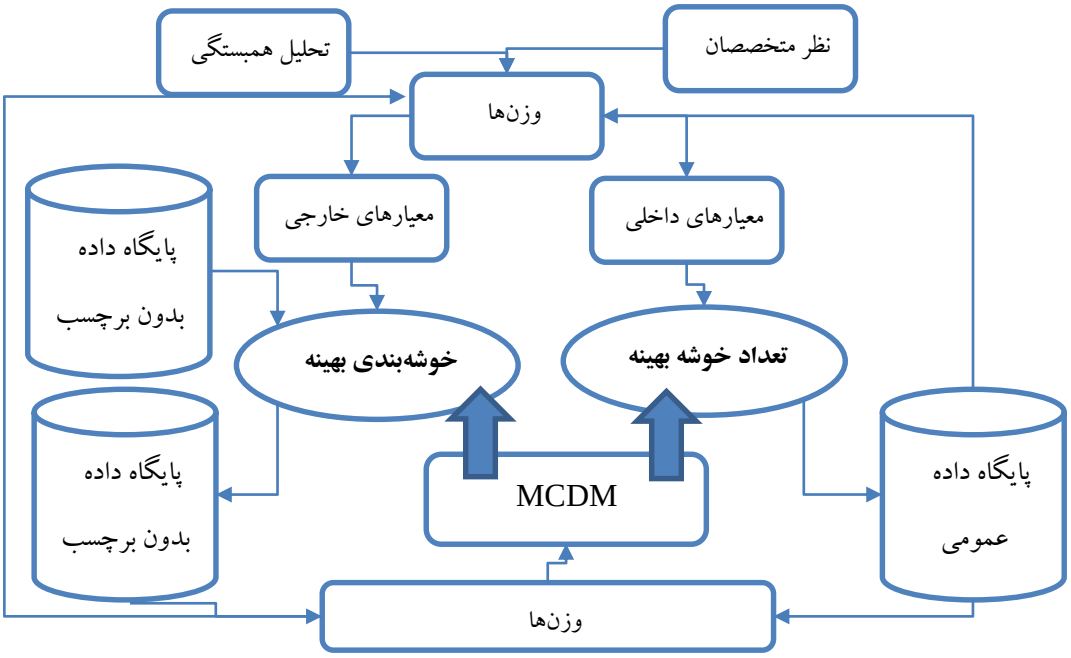
[۱] Crisp

خوشه‌ها و معیارهای اندازه‌گیری AIC, AWE, BIC, CLC, KIC در نظر گرفته‌شده‌اند. تعداد خوشه‌های ۲و۳و۴و۵ به‌عنوان کاندیداها در نظر گرفته‌شده‌اند.

- مرحله دوم: مجموعه‌داده به‌عنوان یک ترکیبی از توزیع‌های نرمال برای تعداد مختلف خوشه‌ها در روش خوشه‌بندی مبتنی بر مدل انتخاب‌شده‌اند. بردارهای میانگین، ماتریس کوواریانس، ترکیب خصوصیات و تابع احتمال توسط الگوریتم EM تخمین زده‌شده است.
- مرحله سوم: مقادیر شاخص‌ها در هر یک از تعداد خوشه‌ها محاسبه می‌شوند.
- مرحله چهارم: ماتریس مقایسه زوجی با استفاده از ماتریس تصمیم تشکیل می‌شود.
- مرحله پنجم: برای هر یک از کاندیداها C-RIV محاسبه شود.
- مرحله ششم: کاندیدایی که بیشترین مقدار C-RIV را دارد بهترین تعداد خوشه‌ها را دارد.

برای آزمون روش از مجموعه‌داده‌هایی استفاده‌شده است که برچسب‌گذاری شده‌اند و نتیجه محاسبه‌شده را با برچسب‌های آن‌ها مقایسه کرده است.

«پنگ» در کار دیگر خود در سال ۲۰۱۲ در یک کار ترکیبی با استفاده از روش MCDM در ابتدا با استفاده از شاخص‌های خارجی بهترین روش خوشه‌بندی را پیداکرده و بعد از پیدا کردن آن با اندازه‌گیری معیارهای ارزیابی داخلی برای آن الگوریتم خوشه‌بندی بهترین تعداد خوشه را محاسبه کرده است (Peng, Zhang, Kou, Li, et al. ۲۰۱۲).



شکل ۲-۱۱ انتخاب بهترین خوشه‌بندی و بهترین تعداد خوشه برای بهترین خوشه‌بندی انتخاب‌شده

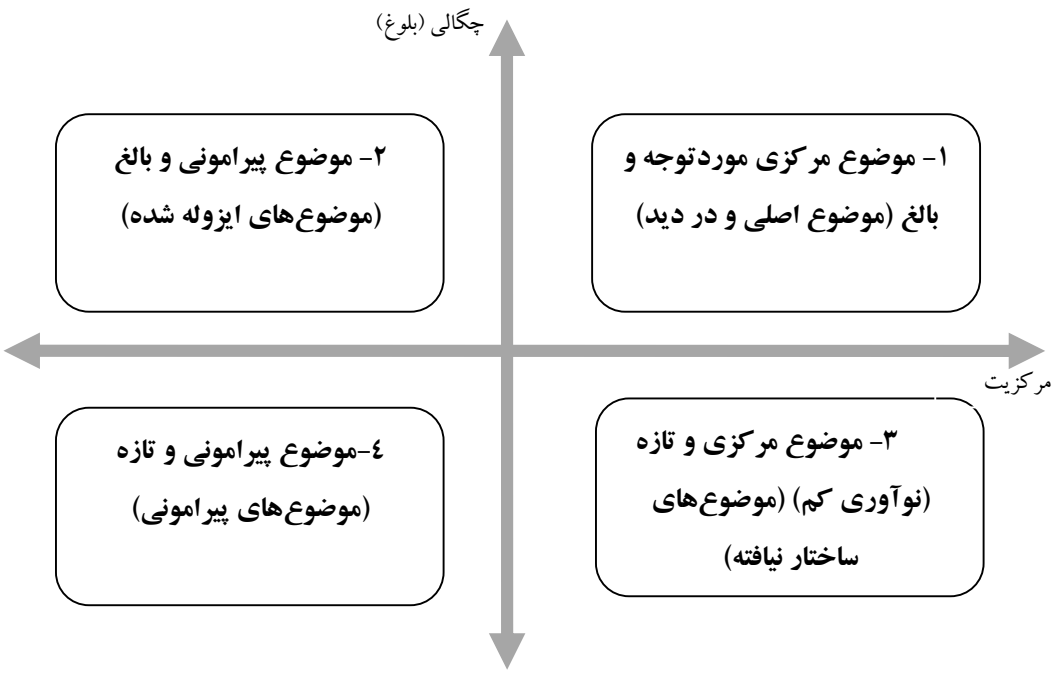
(Peng, Zhang, Kou, Li, et al. ۲۰۱۲)

۲-۱-۹- نمودار راهبردی

با استفاده از نمودار راهبردی تحلیل ساختار و تغییرات روند یک حوزه تحقیقاتی انجام می‌شود. نمودار راهبردی اولین بار در سال ۱۹۸۸ توسط «لا» ارائه شد (Law et al. ۱۹۸۸). این نمودار به‌صورت کلی روابط درون یک حوزه و ارتباط بین حوزه‌ها را نشان می‌دهد. در این نمودار دو بُعد محوریت و چگالی وجود دارد و مکان هر خوشه بر اساس مقدار این دو معیارها در نمودار مشخص می‌شود (Delecroix and Epstein ۲۰۰۴).

هرچه میزان مرکزیت یک خوشه (حوزه) بیشتر باشد نشان‌دهنده این است که آن خوشه مرجعیت بیشتری دارد و با دیگر خوشه‌ها ارتباط بیشتری دارد. میزان مرکزیت یک خوشه را می‌توان با استفاده از ریشه دوم مجموع مربعات ارتباطات خروجی یا میانگین مقادیر چند لینک اول محاسبه کرد. امتیازدهی به خوشه‌ها بر اساس مرکزیت آن‌ها می‌تواند نشان دهد که در کل شبکه تحقیق کدام حوزه‌ها مرکزیت بیشتری دارند (Shen, Li, and Gu ۲۰۱۳).

چگالی یا همبستگی درونی یک خوشه، معیاری است که قدرت روابط کلماتی که خوشه را می‌سازند را نشان می‌دهد. به‌عبارت‌دیگر بیان می‌کند که کلمات به چه میزان باهم در یک مقاله تکرار شده‌اند. طریقه محاسبه میزان چگالی می‌تواند به‌صورت میانگین ارتباطات گره‌های داخل خوشه، میانه ارتباطات گره‌های داخل خوشه، یا مجموع مربعات ارتباطات درونی یک خوشه باشد. هرچه این مقدار بیشتر باشد خوشه بیشتر به بلوغ و تکامل رسیده است. در نمودار راهبردی محور عمودی نشان‌دهنده میزان چگالی (همبستگی درونی یا میزان تازه بودن حوزه) و محور افقی میزان مرکزیت (توجه به حوزه یا در حاشیه نبودن حوزه) است (An and Wu ۲۰۱۱, Shen, Li, and ۲۰۱۳Gu). شکل ۲-۱۲ ساختار اصلی نمودار راهبردی نمایش داده‌شده است.



شکل ۲-۱۲ نمودار راهبردی

در این پژوهش، پس از خوشه‌بندی، مرکزیت و چگالی هر خوشه محاسبه‌شده و برای تحلیل بهتر بر روی نمودار راهبردی ترسیم می‌شوند. اندازه چگالی هر خوشه بر اساس رابطه (۹) و مرکزیت خوشه‌ها بر اساس رابطه (۱۰) محاسبه می‌شود. برای به دست آوردن مرکزیت از میانگین مجموع وزن‌های ارتباطات بین کلیدواژه‌های متعلق به خوشه موردنظر با کلیدواژه‌هایی که متعلق به آن خوشه

نیستند استفاده‌شده است. برای محاسبه چگالی خوشه از میانگین وزن هر کلیدواژه خوشه و مجموعه وزن‌های ارتباطات کلیدواژه‌های همان خوشه استفاده‌شده است.

۲-۲- تحلیل‌روند و پیش‌بینی علم و فناوری

در این بخش به بررسی پیشینه‌پژوهش در خصوص تحلیل‌روند و پیش‌بینی علم و فناوری خواهیم پرداخت. روش‌های تحلیل‌روند استخراج شده و مورد ارزیابی قرار خواهند گرفت.

۲-۲-۱- مقدمه

بررسی روند تحقیقات یک حوزه علمی می‌تواند درک بهتری را برای محققین و سیاست‌گذاران آن حوزه ایجاد نماید تا بتوانند برنامه‌ریزی مناسبی را جهت انجام تحقیقات آتی و تخصیص منابع پژوهشی داشته باشند. یکی از مهم‌ترین رویکردها در تحلیل‌روند تحقیقات در یک حوزه، بررسی اسناد علمی منتشرشده در آن حوزه با استفاده از روش‌های علم‌سنجی و پیمایش اطلاعات و متن اسناد است. بنابراین تحلیل‌روند ابزار مناسب برای محققین و سیاست‌گذاران در اجرای فعالیت‌های آن‌ها است و دقت و جامعیت تحلیل‌روند از اهمیت ویژگی‌ای برخوردار است. از این‌رو انتخاب روشی مناسب برای تحلیل وضعیت فعلی و پیش‌بینی آن حائز اهمیت است.

از سویی دیگر از معنای لغوی اصطلاح پیش‌بینی می‌توان این درک را داشت که منظور تجسم یک موقعیت یا وضعیت در آینده است. شرکت‌های بزرگ و موفق برای جلب رضایت مشتری‌های خود و در پی آن کسب ارزش‌افزوده بیشتر از فناوری‌های برتر استفاده می‌کنند. ازاین‌رو برنامه‌ریزی فناوری امری حیاتی است. در بیشتر کشورهای اروپایی (خصوصاً سازمان توسعه و همکاری اقتصادی) پیش‌بینی فناوری و برنامه‌ریزی فناوری به‌طور رسمی از دهه ۱۹۶۰ آغاز شد.

۲-۲-۲- تعریف و تقسیم‌بندی کلی روش‌های پیش‌بینی فناوری

در پیش‌بینی فناوری، «فناوری» ابزارها، شیوه‌ها و فرایندهایی است که باهدف برآورده کردن نیازهای انسان مورداستفاده قرار می‌گیرد (Martino۱۹۹۳). پیش‌بینی فناوری را به‌طورکلی هرگونه روش هدفمند و نظام‌مند برای پیش‌بینی و فهم در زمینه‌های نرخ رشد فناوری، جهت رشد

فناوری، ویژگی‌های رشد آن و تاثیرات تغییرات فناوری (بخصوص ابداعات، نوآوری‌ها) خوانده‌اند (Firat, Woon, and Madnick ۲۰۰۸). پیش‌بینی فناوری عبارت است از پیش‌بینی تمایلات و روند فناوری مانند جهت رشد و نرخ توسعه فناوری (Jun and Lee ۲۰۱۲, Jun ۲۰۱۱, Porter ۱۹۹۱a).

برای پیش‌بینی فناوری از روش‌های مختلفی استفاده‌شده است. این روش‌ها را می‌توان به گونه‌های مختلفی تقسیم‌بندی کرد و پیشگو می‌تواند بر اساس استفاده از یک یا ترکیب چند روش، پیش‌بینی را انجام دهد. در ادامه سه نوع تقسیم‌بندی مختلف را بیان کرده‌ایم که تقسیم‌بندی آخر تکمیل‌کننده تقسیم‌بندی دوم است. در سال ۱۹۹۳ «ژورف» در کتاب خود (Martino ۱۹۹۳)روش‌های پیش بینی‌فناوری را جدا از ناحیه کاربرد پیش‌بینی و براساس نحوه *به دست* آوردن خروجی سیستم از داده‌های ورودی به چهار دسته تقسیم نمود:

۱. **روش‌های برون‌یابی**[۱]**:** این روش برای پیش‌بینی کردن، الگویی که درگذشته پیداشده است را مورد بسط یا نگاشت به شرایط حاضر قرار می‌دهد. پیشگو یک بازه زمانی را از گذشته انتخاب کرده و یک مورد خاص را موردمطالعه قرار می‌دهد تا به یک الگوی خاص از موردمطالعه برسد. الگوهای معمول تمایلات و یا چرخه‌هایی هستند که درگذشته اتفاق افتاده‌اند. به‌محض اینکه الگویی پیدا شد، سعی می‌کنند تا الگو را برای آینده بسط دهند . مبنای این روش بر این اصل استوار است که اطلاعاتی که از گذشته به دست می‌آیند برای پیش‌بینی آینده کافی است.

۲. **شاخص‌های پیشرو**[۲]**:** در این روش، پیشگو بر مبنای مطالعه اطلاعات یک بازه زمانی، سعی بر پیدا کردن شاخص‌ها و معیارهایی دارد تا از طریق آن‌ها بتواند آینده را مورد پیش‌بینی قرار دهد.

---

[۱] Extrapolation

[۲] Leading indicators

۳. **مدل سببی(علی)**[۱]**:** دو روش قبلی مبتنی بر ارتباط بین گذشته و آینده است. همان‌طور که از نام این روش پیداست رابطه علت و معلول را پیدا می‌کند. این روش می‌تواند روابط را به‌صورت ریاضی به شکل‌های مختلفی بیان کند.

۴. **روش‌های احتمالاتی**[۲]**:** در سه روش گذشته با استفاده از داده‌های داشته شده، فقط یک پیش‌بینی از آینده را به ما می‌دهند. اما در این روش بجای اینکه یک جواب برای آینده *به دست* آید، مجموعه‌ای از پیش‌بینی‌ها با احتمالات مختلف را به ما ارائه می‌دهد.

همان‌طور که در بالا می‌توان ملاحظه کرد، نوع داده‌های ورودی که قرار است بر اساس آن‌ها پیش‌بینی انجام شود در تقسیم‌بندی فوق نقشی نداشته است و در ضمن هیچ‌گونه روش نظری برای پیش‌بینی مورداستفاده قرار نگرفته است. تقسیم‌بندی دیگری که در سال ۱۹۹۱ «پورتر» در کتاب خود Porter) ۱۹۹۱(ارائه کرده است، پیش‌بینی فناوری را بر اساس نحوه پیش‌بینی و داده‌های ورودی برای پیش‌بینی به‌صورت زیر تقسیم‌بندی کرده است:

۱. **پایش**[۳]: پایش به معنای تهیه تصویری از فناوری خاص از محیط با استفاده از اطلاعات در دسترس است. این اطلاعات امکان دارد به فناوری خاصی مربوط باشد. پایش را فرایند شناسایی نشانه‌های ممکن در بخش‌های علمی، اقتصادی، مدیریتی، سیاسی یا نظامی که ممکن است منجر به پیشرفت‌های احتمالی در فناوری نیز شود، تعریف می‌کنند .این روش امکان آگاهی از تغییراتی که می‌تواند در نفوذ و پذیرش فناوری‌ها در بازار تأثیر بگذارد را به ما می‌دهد (Madnick در Madnick .(۲۰۰۷Phillips, Heidrick, and Potter) بیان کرده (۲۰۰۸Madnick et al. که پایش محیط می‌تواند به‌عنوان ورودی اصلی برای تحقیقات آینده باشد. به‌طور کل مشاهده و دنبال کردن گسترش فناوری که در محیط پیرامون در حال رخداد است و به موضوع موردمطالعه مربوط است را پایش گویند (Kim۲۰۰۷).

---

[۱] Causal models

[۲] Probabilistic methods

[۳] Monitoring

۲. **نظرات متخصصان**[۱]: گفته می‌شود که افراد دارای تخصص قابلیت پیش‌بینی‌های بهتری درزمینه‌ی تخصصی خوددارند. که البته لازمه آن شناسایی و استفاده از افراد توانا درزمینه‌ی موردنظر است. چنانچه این شرایط فراهم نباشد، این روش نمی‌تواند مناسب باشد. همچنین استفاده از نظرات کارشناسان تضمینی برای پیش‌بینی موفق نیست. درهرحال یکی از مهم‌ترین روش‌های مورداستفاده است (Van Der Heijden۲۰۰۰).

۳. **برون‌یابی روند**[۲]: این روش بر این فرض استوار است که آینده بر اساس رویدادها، روندها و الگوهای پیشرفتی درگذشته قابل‌رؤیت هستند سنجیده می‌شود. با این تصور که تغییرات گذشته در آینده ادامه خواهد یافت و این عوامل تغییر چندانی نمی‌کنند و بر این اساس می‌توان آینده را تا حدودی پیش‌بینی کرد.

تحلیل‌روند از فنون ریاضی و آمار بهره می‌گیرد تا اطلاعات سری‌های زمانی را به آینده تعمیم دهد. فنون تحلیل‌روند ازلحاظ پیچیدگی متفاوت‌اند این فنون از شکل ساده رگرسیون، تطبیق با منحنی[۳] تا شیوه‌فنی Box-Jenkins را شامل می‌شود (Levary and Han۱۹۹۵). فناوری مراحل گوناگونی در چرخه زندگی خود طی می‌کند که شامل به دنیا آمدن، رشد و تکامل، بلوغ و زوال فناوری است. با منحنی‌های رشد و زوال می‌توان نقطه‌هایی که رشد فناوری یا زوال آن به حداکثر می‌رسد را تشخیص داد. بنابراین روشی کمی است که از داده‌های گذشته (تاریخی) با توابع ریاضی آینده را پیش‌بینی می‌کنند (Bengisu and Nekhili۲۰۰۶).

۴. **مدل‌سازی**[۴]: مدل کردن یک ارائه ساده از پویایی ساختار[۵] بعضی بخش‌های دنیای واقعی است. مدل می‌تواند رفتار آینده یک سیستم پیچیده را به‌سادگی نشان دهد که این کار به‌وسیله مجزا سازی ویژگی‌های مهم سیستم از جزئیات غیرضروری است. یکی از روش‌هایی که روابط

---

۱ Expert Opinion

۲ Trend Extrapolation

۳ Curve fitting

۴ Modeling

۵ Structure dynamics

بین رویدادها را در نظر می‌گیرد، روش مدل‌سازی است. این روش یا بر پایه کامپیوتر (مانند شبیه‌سازی) و یا بر پایه قضاوت است. در بعضی مواقع مدل‌ها علاوه بر اینکه رفتار آینده را به ما می‌دهند، رابطه بین این پیش‌بینی و متغیرهای آن را نیز نشان می‌دهند (Daim et al. ۲۰۰۶).

۵. **سناریوها**[۱]: پیش‌بینی به روش سناریو شامل تشخیص رویدادها و یا رخدادهای کلیدی، که ممکن است در تکامل آینده فناوری نقش داشته باشند، است (McDowall and Eames ۲۰۰۶). سناریوهایی که در حوزه پیش‌بینی فناوری بکار برده می‌شوند، مفاهیم مختلف فناوری‌های آینده را توصیف کرده و گزینه‌های فناوری آینده را نشان می‌دهند. سناریوها زمانی مفیدند که اطلاعات در دوره‌های زمانی گذشته در دسترس نباشد یا متخصصان درزمینهٔ مربوط ضعیف بوده یا وجود نداشته باشند و هیچ پایه محکمی برای ایجاد مدل وجود نباشد. سناریوها عبارت‌اند از یک سری تصاویر ترسیمی از آینده (با حوادث آینده) که ما را از زمان حال به آینده رهنمود می‌کنند. هر مجموعه از سناریوها، شامل احتمالاتی منطقی در مورد ابعاد خاصی از آینده است.

«فیرات» و همکاران در (Firat, Woon, and Madnick ۲۰۰۸) تقسیم‌بندی فوق را بسط داده و به زیرمجموعه‌هایی از آن‌ها نیز اشاره کرده‌اند که ما به خلاصه‌ای از آن‌ها در زیر اشاره می‌کنیم :

۱. **نظرات متخصصان:** شامل روش‌های زیر است:
   a. دلفی
   b. گروه‌های کانونی[۲]
   c. مصاحبه
   d. روش‌های مشارکتی
۲. **تحلیل‌روند:** شامل روش‌های زیر است:
   a. برون‌یابی روند
   b. تحلیل تأثیرات روند

---

[۱] Scenario

[۲] Focus Groups

۳. **پایش و نظارت بر اطلاعات:** شامل روش‌های زیر است:

a. پایش (پیمایش محیط و فناوری‌ها)

b. روش کتابشناختی که شامل متن‌کاوی و تحلیل منابع علمی و ثبت اختراع است

٤. **روش‌های آماری:** شامل روش‌های زیر است:

a. تحلیل همبستگی

b. تحلیل ریسک

c. روش کتابشناختی که شامل متن‌کاوی و تحلیل منابع علمی و ثبت اختراع است

۵. **مدل کردن و شبیه‌سازی**: شامل روش‌های زیر است:

a. مدل کردن عاملی

b. مدل کردن نفوذ

c. ارزیابی فناوری

۶. **سناریو:** شامل روش‌های زیر است:

a. سناریو و مدیریت سناریو

b. شبیه‌سازی سناریو و سناریوهای تعاملی

۷. **روش‌های ارزش‌گذاری، تصمیم‌گیری، و اقتصادی:** محبوب‌ترین روش در این بخش رهیافت درخت ارتباط[۱] است. در این روش اهداف فناوری به بخش‌ها و سطوح پایین‌تری شکسته می‌شود. در این روش ساختار سلسله مراتبی توسعه فناوری شناخته می‌شوند. احتمال دسترسی به هر یک از سطوح *به دست می‌آید*. احتمالات برای *به دست* آوردن فناوری هدف مورداستفاده قرار می‌گیرند (Levary and Han۱۹۹۵).

a. درخت ارتباط

b. تجزیه‌وتحلیل هزینه و سود

[۱] Relevance tree approach

۸. **روش‌های توصیفی:** قیاس‌ها[1] از روش‌های محبوب مورداستفاده در این بخش هستند. استفاده از قیاس در پیش‌بینی شامل مقایسه نظام‌مند فناوری موردنظر با فناوری‌های قبلی است که به‌صورت مشابه در بسیاری از جنبه‌ها در حال کار کردن هستند. یکی از کمبودهای قیاس این است که فرض کرده انسان‌ها به‌صورت معمول قبل رفتار می‌کنند و آن‌ها را می‌توان با قرار دادن در موقعیت‌های مشابه، رفتار مشابهی از آن‌ها دریافت کرد.

a. قیاس
b. فهرست بندی برای شناسایی تأثیرگذارها
c. ارزیابی تأثیرات اجتماعی

۹. **خلاقیت:** شامل روش‌های زیر است:

a. طوفان فکری
b. کارگاه‌های خلاقیت
c. تحلیل علمی تخیلی
d. تولید چشم‌انداز

بعضی از این روش‌ها بر اساس قضاوت و نظر متخصصان صورت می‌گیرد. داشتن پیش‌فرض‌ها، تعصبات و تمایلات شخصی در استفاده از نظرات متخصصان می‌تواند باعث اشتباه در آینده‌نگری شود و برنامه‌ریزی‌های بلندمدت و میان‌مدت را به خطر اندازد. روش سناریونویسی نیز مانند نظر متخصصان یک روش کیفی است و معمولاً برای مواردی مورداستفاده قرار می‌گیرد که اطلاعاتی از دوره‌های زمانی گذشته در رابطه با گذشته یک فناوری در دسترس نباشد یا متخصصان درزمینه مربوط ضعیف بوده یا وجود نداشته باشند. احتمال درست درآمدن پیش‌بینی از طریق سناریو کم است (Martino ۲۰۰۳). در مدل‌سازی علاوه بر پیچیده بودن و زمان‌بر بودن ساخت یک مدل برای فناوری یک کشور، این احتمال وجود دارد که تمام حالات و متغیرهای مرتبط شناخته‌نشده باشند

[1] Analogies

چه رسد به ارتباط بین این حالات و متغیرها و درنتیجه مدل درجاهای حساس به‌درستی عمل نکند (Baharudin, Lee, and Khan ۲۰۱۰).

استفاده از داده‌های عددی دیگر پیش قضاوت‌ها و تعصبات نظرات اشخاص را ندارد و می‌تواند دقت در پیش‌بینی را زیاد کرده و آن را مستند و مکتوب کند. روش‌های آماری این ویژگی‌ها رادارند. اما درزمانی که از این روش‌ها استفاده می‌کنیم امکان دارد شرایطی برقرار باشد که در زمان گذشته وجود نداشته باشد و می‌تواند در پیش‌بینی آینده ما خدشه ایجاد کند. ازآنجاکه برون‌یابی روند نیز از داده‌های تاریخی استفاده می‌کند، این مشکل را در خود دارد. از طرفی دیگر داده‌ها باید به‌صورت کاملاً عددی در اختیار این سیستم گذاشته شوند و قابلیت استفاده از اطلاعاتی که به‌صورت نهفته در اسناد و مدارک وجود دارند را این دو روش نمی‌توانند در نظر بگیرند. نقصان در داده‌های عددی و روش محاسباتی نامتناسب را می‌توان از معایب این شیوه دانست. ازجمله به انتخاب منحنی رشدی که متناسب با شرایط فعلی نباشد، می‌توان اشاره کرد. انتخاب برای استفاده از منابع متناسب با فناوری هدف و سطح رشد آن فناوری، در روش پایش از مشکلات این روش است. برای این منظور می‌بایست از نظرات متخصصان استفاده کرد.

۲-۲-۲-۱- پیش‌بینی آینده با استفاده از روش‌های کتابشناختی

همان‌طور که ملاحظه کردید یکی از روش‌های پیش‌بینی فناوری که در دو مجموعه پایش و تحلیل‌های آماری است، روش کتابشناختی است. کتابشناختی مجموعه ابزاری است که برای تحلیل داده‌های متنی مورداستفاده قرار می‌گیرد. «نورآون» و همکاران در (Norton ۲۰۰۰) کتابشناختی را معیاری برای ارزیابی اطلاعات و متن دانسته‌اند این اطلاعات شامل نام نویسنده، محل نگارش، ارجاعات به آن متن، کلمات‌کلیدی است. کتابشناختی روشی برای ارزیابی، توصیف و نظارت بر روی انتشارات است که می‌تواند زمینه‌های مربوطه را مورد تحلیل قرار داده و کیفیت و تمرکز تحقیقات یک موسسه بر روی یک موضوع خاص را موردبررسی قرار دهد.

فناوری کاوی به روش‌هایی برای استخراج اطلاعات معنی‌دار از پایگاه داده‌های متنی و مرتبط با یک فناوری خاص گفته می‌شود (Porter and Cunningham ۲۰۰۵). روش‌های متعددی نیز برای پیدا کردن تمایلات و روند رشد فناوری ارائه‌شده است (Rahayu and Hasibuan ۲۰۰۶ ,

۲۰۰۵, Kobayashi et al. ۲۰۰۹Woon, Henschel, and Madnick). برخی از این روش‌ها با استفاده از روش‌های کتابشناختی بر روی مقالات منتشرشده و اسناد اختراعات ثبت‌شده تحلیل‌های پیش‌بینی خود را انجام داده‌اند (Fattori, Pedrazzi, and Turra ۲۰۱۲Jun, Park, and Jang , , Tseng, Lin, and Lin ۲۰۰۸, Kim, Suh, and Park ۲۰۰۹, Lee, Yoon, and Park ۲۰۰۳ ۲۰۱۲, Yoon and Kim ۲۰۰۸b, Yoon, Phaal, and Probert ۲۰۱۱, Jun ۲۰۰۷). در هرکدام از روش‌ها سعی شده که اسناد بر اساس معیارهایی سازمان‌دهی شوند. برخی از قوانین همبستگی (Ghani and Fano ۲۰۰۲, Pierre ۲۰۰۲) و برخی دیگر با استفاده از روش‌های کلاس‌بندی (۲۰۰۴, Liu et al. ۲۰۰۲Ghanem et al.) و دسته‌بندی (Hotho, Staab, and Stumme ۲۰۰۳ , ۲۰۰۲Beil, Ester, and Xu)برای سازمان‌دهی اسناد استفاده کردند.

ازآنجاکه تحلیل‌های بعدی بسیار وابسته به چگونگی سازمان‌دهی اسناد است این سازمان‌دهی بسیار حائز اهمیت است. بنابراین استفاده از یک روش مناسب برای سازمان‌دهی می‌بایست لازمه تحلیل است. اما برای سازمان‌دهی نیازمند این هستیم که بتوانیم ویژگی‌هایی از این اسناد استخراج‌کنیم تا بر اساس این ویژگی‌ها اسناد را مرتب کنیم. روش‌های گوناگونی برای انتخاب ویژگی‌ها از متون ارائه‌شده است (Lee and ۱۹۹۸, Liu and Motoda ۲۰۰۵Wang and Wang , , ۲۰۰۹, Chen et al. ۲۰۰۵, Yan et al. ۲۰۰۵, Manomaisupat and Ahmad ۲۰۰۶Chen , ۲۰۰۹, Aghdam, Ghasem-Aghaee, and Basiri ۲۰۰۹Ogura, Amano, and Kondo).

۲-۲-۳- تعاریف، روش‌ها و شاخص‌های اندازه‌گیری موضوع‌های علمی و فناوری (نوظهور)

ازآنجاکه برای تحلیل و پیش‌بینی ابتدا باید حوزه موردنظر را شناخت، لازم است که در ابتدا به کشف موضوعات و حتی موضوعات نوظهوری در مستندات و اطلاعات کتابشناختی پرداخته شود. محققان بیان می‌کنند که هرچند که مفاهیم در این حوزه (فناوری‌ها و موضوع‌های نوظهور) وجود دارند، اما ویژگی‌های اصلی و بنیادی آن هنوز مبهم است و ارتباط بین تعریف و رهیافت رسیدن به آن بسیار سست و شکننده است (۲۰۱۸Wang) و (۲۰۱۵Rotolo, Hicks, and Martin). علاوه بر این محققین بیان کرده‌اند که مفهوم و ویژگی‌های نوظهور بودن به‌طور کامل و جامع تعریف‌نشده

است و ارتباط بین مفهوم و شاخص‌های پیشنهادی بسیار مبهم است (Wang۲۰۱۸) و (Small, Boyack, and Klavans۲۰۱۴) و (Cozzens et al.۲۰۱۰).

به‌صورت کلی بر اساس مطالعات انجام‌شده کشف موضوع‌های جدید در دو مرحله انجام می‌گیرد: ساختن خوشه‌هایی از خروجی‌های علمی (مبتنی بر اطلاعات کتابشناختی از قبلی کلیدواژه‌ها، عنوان، چکیده، یا ارجاعات)، و در مرحله دوم استفاده و یا ساخت معیارهایی برای تشخیص موضوع‌ها. روش‌هایی برای *به دست* آوردن موضوع‌های تحقیقاتی ارائه‌شده است، بر اساس نظر «سنگ» در (Tu and Seng۲۰۱۲) (۲۰۱۲)روش‌های پیشین در خصوص کشف موضوع به سه دسته تقسیم می‌شوند:

- روش‌های مبتنی بر متن‌کاوی و داده‌کاوی
- روش‌های مبتنی بر اندازه‌گیری واژه‌ها (ویژگی‌ها) و خط زمانی
- روش‌های تحلیلی ترکیبی و تحلیل لینک

«ونگ» در پژوهش خود در سال ۲۰۱۸ روش‌های کشف رشته‌های نوظهور را عموماً بر سه دسته تقسیم کرده است (Wang۲۰۱۸):

- **روش‌های مبتنی بر هم‌استنادی:** برای مثال «اسمال» در ۲۰۱۴ روشی مبتنی بر خوشه‌بندی هم‌استنادی و استناد مستقیم در پایگاه داده‌های بزرگ برای تشخیص موضوع‌های نوظهور ارائه کرده است (Small, Boyack, and Klavans۲۰۱۴).
- **روش‌های مبتنی بر متن:** (در بخش‌های قبلی و بعدی مفصل در این رابطه صحبت شده است).
- **روش‌های ترکیبی:** برای مثال «چن» در سال ۲۰۰۶ نیز با استفاده از هم‌استنادی و روابط بین کلمات موضوع‌های جدید را کشف کرده است.

لازم به ذکر است که روش‌های مدل‌سازی موضوعی نیز در دسته دیگری از این‌ها می‌توانند قرار بگیرند برای مثال در(Tu and Seng۲۰۱۲) و (Jo, Lagoze, and Giles۲۰۰۷) و (Wang ۲۰۱۸) از روش‌های مدل‌سازی موضوعی استفاده‌شده است که در دسته‌بندی ایشان ذکر نشده است.

«ماریو» در سال ۲۰۱۰ با استفاده از تشخیص کلمات نوظهور (کلماتی که در یک بازه زمانی زیاد مورداستفاده قرار می‌گیرند ولی قبل از آن کم بوده یا نبودند)، تشخیص موضوع‌های نوظهور (مجموعه کلماتی که با کلمات نوظهور *هم‌رخداد* بوده‌اند) را انجام داده است (Cataldi, Di Caro, and Schifanella ۲۰۱۰). ویژگی‌های فناوری‌های نوظهور که در (Cozzens et al. ۲۰۱۰) توسط «کوزنس» در سال (۲۰۱۰) ارائه شد:

- سرعت‌بالای رشد (در بازه زمانی اخیر)
- تغییر به شکل جدید
- پتانسیل بازار و اقتصاد
- افزایش پایگاه علمی (مستندات علمی)

ویژگی‌های نوظهور در (Glänzel and Thijs ۲۰۱۱) توسط «گلنزل» در سال (۲۰۱۱) بیان شد:

- خوشه‌هایی که رشد زیادی دارند
- خوشه‌های جدیدی که تشکیل‌شده‌اند (این خوشه‌ها حتماً از ترکیب خوشه‌های دیگر به وجود آمده‌اند).
- خوشه‌هایی که موضوع آن‌ها عوض‌شده است

ازنظر (Tu and Seng ۲۰۱۲) (۲۰۱۲) برای تشخیص یک فناوری نوظهور دو پارامتر لازم است:

- شاخص جدیدی بودن
- شاخص میزان انتشار

البته نکته قابل‌توجهی که در سال ۲۰۱۳ توسط «موهاند هلاوه» در (Halaweh ۲۰۱۳) در خصوص ویژگی برخی از فناوری‌ها مانند آی‌تی بیان کرده است محدودیت در برخی کشورهای خاص، نبود و ضعف در تحقیقات، نگرانی‌های اعتقادی و تأثیرات پیش‌بینی‌نشده اجتماعی است.

در خصوص ویژگی‌های نوظهور بودن تنها بر سر دو ویژگی برای نوظهور بودن به‌صورت کلی توافق شده است (۲۰۱۴) (Small, Boyack, and Klavans ۲۰۱۴):

- جدید بودن
- رشد سریع

ویژگی‌های فناوری‌های نوظهور که توسط «روتول» با جمع‌بندی کارهای گذشته در زمینه‌های تحقیقاتی مختلف به دست آورده و در (Rotolo, Hicks, and Martin ۲۰۱۵) در سال (۲۰۱۵) ارائه‌شده است به‌قرار زیر است:

- به‌صورت بسیار زیاد جدید (جدید بودن صرفاً به معنی روش جدید نیست، بلکه کاربرد جدید هم محسوب می‌شود) (۲ مطالعه از ۱۲ مطالعه)
- سرعت رشد بالا (سه مطالعه از ۱۲ مطالعه).
- انسجام: وابسته به ویژگی‌های درونی یک گروه است (شامل باهم آمدن، ترکیب با همدیگر، و یکپارچه بودن و انطباق است) (۴ مطالعه از ۱۲ مطالعه).
- تأثیر مهم بر روی بخش‌های زیاد (برای مثال خلق صنعت جدید، تغییر صنعت فعلی) (۹ مطالعه از ۱۲ مطالعه).
- غیرقطعی (به خاطر شرایط اجتماعی و اقتصادی) و مبهم (به خاطر شرایط اجتماعی و اقتصادی) و قابلیت تغییر مسیر فعلی (۷ مطالعه از ۱۲ مطالعه).

شاخص‌هایی نیز که تابه‌حال برای روش‌های تشخیص فناوری‌های نوظهور ارائه‌شده است نیز به دسته‌های زیر قابل‌تقسیم شدن است (۲۰۱۸) (Wang۲۰۱۸):

- برحسب تعداد مستندات علمی چاپ‌شده در سال: برای مثال در (Small, Boyack, and Klavans۲۰۱۴)
- تشخیص سند مرکزی[۱] در خوشه و دنبال کردن آن تا کشف موضوع‌های نوظهور
- روابط هم‌استنادی که با خارج از خوشه موردنظر برقرار می‌شود.
- شمارش کلیدواژه‌ها یا کلمات جدید برای تشخیص موضوع‌های جدید.
- شمارش و جابجایی میزان نویسندگان در مقالات.
- روش‌های ترکیبی از موارد فوق.

با استفاده از اطلاعات کتابشناختی این پارامترها قابل‌اندازه‌گیری هستند اما این دو ویژگی (سرعت شد و جدید بودن) تنها بخشی از ویژگی‌های نوظهوری را نشان می‌دهند و ویژگی‌ها دیگر

[۱] Core Document

از قلم افتاده می‌شوند (Rotolo, Hicks, and Martin۲۰۱۵). محدودیت‌های دیگری که در این خصوص به وجود می‌آید تعداد محدود مقالات چاپ‌شده یا پارس‌ها برای تحلیل و در ثانی محدود بودن کلیدواژه‌های استخراج‌شده از آن‌ها است. بنابراین با استفاده از خروجی‌هایی که از این مطالعات محدود و نتایج *به‌دست‌آمده* نمی‌توان به مواردی که برای پیش‌بینی سازمان‌ها موردنظر قرار می‌گیرد دست یافت و می‌بایست پایگاه‌های داده بزرگی وجود داشته باشد.

برای الگوهای علم نیز شرایط به وجود آمده در محیط علمی بعد از به وجود آمدن موضوع جدید قابل‌مشاهده است (Wang۲۰۱۸). هم‌رخدادی کلمات می‌تواند ساختار فکری و زمینه تحقیقاتی یک حوزه را آشکار کند(Hu and Zhang۲۰۱۵). تعاریف ارائه‌شده در خصوص فناوری‌های نوظهور به‌قرار زیر هستند:

به‌صورت کلی، یک فناوری‌ای جدیدی که نسبتاً رشد سریعی دارد با میزانی از انسجام در طول زمان به همراه است که قابلیت اعمال و اثر روی دامنه‌ای از اقتصاد و اجتماع داشته باشد. هرچند که این امر وابسته به آینده است و در فاز نوظهوری بسیار غیرقطعی و مبهم است (Rotolo, Hicks, and Martin۲۰۱۵)

در (Tu and Seng۲۰۱۲) یک‌روند موضوع نوظهور را به‌صورت یک **موضوع** در حال **رشد** موردعلاقه یا کاربردی روی یک بازه زمانی معرفی کرده است. به عبارتی یک موضوع نوظهور یک موضوع تحقیقاتی است که مهم است ولی هنوز به موضوع داغ (بلوغ) تبدیل نشده است (Tu and Seng۲۰۱۲).

تعریف ارائه‌شده در خصوص موضوع‌های نوظهور در (Wang۲۰۱۸) به این صورت است: یک موضوع نوظهور موضوعی است که بسیار جدید و سرعت تحقیق در آن نسبتاً بسیار سریع با درجه‌ای از انسجام است و تأثیر قابل‌توجهی نیز در محیط علمی بگذارد. بنابراین ویژگی‌های یک موضوع نوظهور به‌قرار زیر است:

- **بسیار جدید:** برای محاسبه این معیار از میزان تغییرات مقالات چاپ‌شده آن خوشه در یک بازه زمانی کوچک به روی مقالات چاپ‌شده در آن سال استفاده‌شده است **(شتاب)** برای اینکه میزان مقالات می‌تواند به‌صورت تصادفی بالا و پایین برود این محاسبه با میانگین

گرفتن سال بعدی و قبلی اسموت شده است $Rate_{i,t,delta} \geq Rate_{min}$.

- **نسبتاً رشد سریع:** برای این پارامتر موضوع‌ها در ابتدا باید باگذشت زمان میزان مقالات چاپ‌شده در آن‌ها افزایش زیادی پیدا کنند. یعنی در فاز ابتدایی بسیار جدید باشد (بسیار کم چاپ‌شده باشد) و از یک آستانه‌ای کمتر باشند $P_{i,j} \leq P_{min}$
- **انسجام:** میزان انسجام به این صورت بیان‌شده که اگر تعداد استنادهای دریافت شده به یک خوشه از یک مقداری کمتر باشد (مثلاً از تعداد مقالات آن خوشه)، بنابراین آن مقالات اتصال کمی دارند و انسجام کافی را ندارند. $H_i \leq H_{minimum}$
- **تأثیر علمی:** در این پارامتر میزان استنادهایی که به مقالات یک خوشه انجام‌شده است محاسبه گردیده است. درصورتی‌که میزان استنادات به آن خوشه (موضوع علمی) از اندازه‌ای بیشتر باشد آن موضوع تأثیر علمی قابل‌ملاحظه‌ای دارد $Citaione_{i,t,delta} \geq Citation_{\min}$

از سویی دیگر «استفان کارلی» درنتیجه پژوهش‌هایی که برای پروژه FUSE [1] آمریکا که در سال ۲۰۱۸ منتشرشده بیان می‌کند که نوظهور بودن (علم، فناوری، و نوآوری) شامل چهار جنبه است (Carley et al. ۲۰۱۸):

- **تازگی** – برای تازگی باید یک تعریف مشخص ارائه شود.
- **دوام و ماندگاری** – رفتار این شاخص وابسته به دامنه و اندازه مفهوم و حوزه دارد. دوام و ماندگاری می‌تواند منبع یک نویز یا ظهور جدید باشد. این معیار باید آستانه‌ای برای نشان دادن دوام خود داشته باشد.
- **همراهی و اجتماع** – به صورتی که مجموعه‌ای از افراد به آن موضوع علاقه‌مند باشند و آن را دنبال کنند.
- **رشد** – رشد در طول زمان شامل سه بعد می‌تواند باشد: رشد مفهومی و رشد در دیگر موضوع‌ها، و رشد در اجتماع (مثلاً نویسندگان).

[1] Foresight and Understanding from Scientific Exposition (FUSE)

بر اساس مطالعه «روتر» در سال ۲۰۱۵ که بر روی بیش از ۱۸۳ مقاله چاپ‌شده در اسکوپوس انجام‌شده است، به‌طورکلی روش‌هایی که برای پیدا کردن و تشخیص علم و فناوری‌های جدید بوده است به ۵ دسته تقسیم می‌شوند (Rotolo, Hicks, and Martin۲۰۱۵):

۴- روش‌های تحلیل‌روند و تحلیل‌شاخص‌ها که بر روی تعداد اسناد صورت پذیرفته است (۷۵درصد بر روی مقالات ۱۷درصد روی اختراعات ثبت‌شده).

۵- تحلیل هم‌استنادی‌ها که بر روی روابط بین استنادهای مقاله‌ها انجام می‌شود.

۶- تحلیل هم‌رخدادی کلمات که بر روی متن اسناد انجام‌شده است (۷۱درصد بر روی مقالات ۲۹درصد روی پتنت‌ها).

۷- مطالعه بر روی نقشه‌های علمی که موقعیت یک سند (موضوع یا مجموعه‌ای از اسناد) را در یک ساختار گسترده‌تر نشان می‌دهد.

۸- روش‌های ترکیبی که از دو یا چند روش فوق استفاده می‌کنند (۷۱درصد بر روی مقالات ۲۹درصد روی پتنت‌ها).

برای مشاهده کارهای انجام‌شده در هرکدام از دسته‌بندی‌های فوق به جدول ۵ از مقاله (Rotolo, Hicks, and Martin۲۰۱۵) مراجعه کنید. لازم به ذکر است که رفتار نوظهورها بسیار وابسته به حوزه‌ای است که در حال تحلیل است (.Carley et al۲۰۱۷) و (.Carley et al۲۰۱۸).

۲-۲-۴- کشف دانش و موضوع[۱]

هدف از کشف و تعقیب موضوع توسعه فناوری: جستجو، مرتب‌سازی و ساختاربندی اخبار و مطالب متنی است که از رسانه‌های عمومی منتشر می‌شوند. یکی از مفاهیم مهم در TDT سامانه‌های عملیاتی هستند که داده‌ها را به‌صورت پیوسته زمانی که جمع‌آوری می‌شوند به‌صورت برخط مورد پردازش قرار می‌دهند. بیشترین تحقیقات گذشته بر روی بازیابی اطلاعات داده‌های ایستا بوده است. دومین مفهوم مهم در TDT یک رخداد است. در TDT مفهوم موضوع عبارت است از: یک رویداد یا فعالیت منحصربه‌فرد همراه با تمام وقایع و فعالیت‌هایی که به‌طور مستقیم با آن مرتبط هستند.

[۱] Topic Detection

سیستم TDT می‌تواند شامل ۵ قسمت مجزا باشد که عبارت‌اند از: سیستم دنبال کردن موضوع، تشخیص لینک، تشخیص موضوع، تشخیص اولین داستان، بخش‌بندی داستان‌ها (Fiscus and Doddington۲۰۰۲).

بعد از کشف موضوع از کاربردهای TDT استفاده از موضوع‌های کشف‌شده در تحلیل‌روند، توصیف اسناد، هدایت اطلاعات[۱] است (Tu and Seng۲۰۱۲) و (Jo, Lagoze, and Giles ۲۰۰۷).

همان‌طور که می‌دانید رشد تولیدات علمی در سراسر جهان بسیار چشمگیر است. تجزیه‌وتحلیل این تولیدات امری دشوار اما ضروری است. انجام هر نوع فعالیت علمی و عملی در یک حوزه خاص نیازمند آن است تا درک و فهم مناسبی از پژوهش‌های انجام‌شده در آن حوزه و روند تحقیقات آن حاصل شود. به‌طورمعمول سیاست‌گذاران، محققین و پژوهشگران در ابتدای شروع فعالیت خود در یک حوزه علمی، بامطالعه آثار پژوهشگران گذشته سعی بر درک آن حوزه دارند که بسته به میزان و عمق مطالعه، درک نسبی از چهارچوب کلی و وضعیت علمی جاری پیدا خواهد شد. در راستای این درک، پژوهشگران می‌توانند آخرین وضعیت پژوهشی علمی آن حوزه را فهمیده و با در نظر گرفتن روند حرکت آن، یک موضوع مناسب انتخاب کنند. سیاست‌گذار نیز با درک مناسب و دریافت نقشه علمی از وضعیت علمی و فناوری مربوطه می‌تواند با دیدی باز آینده علمی آن حوزه را ترسیم کرده و در خصوص آن سیاست‌های لازم را اتخاذ کند و سرمایه‌گذاری‌های لازم را انجام دهد.

یکی از روش‌های کشف دانش استفاده از تحلیل هم‌رخدادی کلمات است. با استفاده از روش هم‌رخدادی و استفاده از شاخص‌هایی مانند مرکزیت و چگالی قابلیت ارزیابی حوزه‌های علمی در طول زمان وجود دارد (He۱۹۹۹). با استفاده از هم‌رخدادی علاوه بر پیدا کردن سلسله‌مراتب نواحی تحقیقاتی می‌توانیم زمینه‌های تحقیقاتی بالقوه ولی جزئی را تحقیقات را پیدا کنیم. شاخص‌های

[۱] Information navigation

شمول[1] و مجاورت[2] برای این منظور ارائه‌شده‌اند. شاخص شمول برای پیدا کردن سلسله‌مراتب نواحی موضوعی مورداستفاده قرار می‌گیرد و بر اساس شمارش تعداد اسنادی که در آن دو کلیدواژه وجود دارند بر روی حداقل میزان تکرار هرکدام از کلیدواژه‌ها در کل اسناد محاسبه می‌شود. شاخص به‌صورت رابطه زیر محاسبه خواهد شد و مقدار آن همواره بین صفر و یک است:

$$I_{ij} = \frac{C_{ij}}{\min(C_i, C_j)} \qquad (۵۳-۲)$$

در محاسبه میزان قدرت بین گره‌ها در برخی موارد از شاخص همبستگی[3] استفاده‌شده است که طریقه محاسبه آن در زیر آمده است:

$$E_{ij} = \frac{C_{ij}^{۲}}{\mathrm{C_i} * C_j} \qquad (۵۴-۳)$$

برای تشخیص موضوع‌های جدید «شیوا پراساد» در سال ۲۰۱۱ روشی با استفاده از دسته‌بندی KNN ارائه کرده است. در این روش اسناد جدید مشخص هستند و درصورتی‌که اسناد دیگر کمتر فاصله‌ای از آن‌ها قرار داشته باشند اسناد جدید محسوب می‌شوند (وزن دهی اسناد بر اساس TF.IDF انجام‌شده است). با استفاده از روش خوشه‌بندی کا-میانه و استفاده از دیکشنری موضوع‌های جدید را به دست می‌آورد. ازآنجاکه داده‌های برچسب‌گذاری شده استفاده شده است روش ارزیابی با استفاده از معیارهای بازیابی اطلاعات ازجمله معیار F۱ است (Kasiviswanathan et al. ۲۰۱۱).

«جلالی‌منش» در سال ۲۰۱۲ روشی برای کشف دانش از پایگاه‌داده‌های علمی با استفاده از هم‌رخدادی کلمات و تحلیل شبکه‌های اجتماعی ارائه کرده است. تحلیل ایشان با استفاده از ۶۵۰ عنوان پارسا حوزه مهندسی صنایع بوده است. ایشان بعد از حذف کلمات‌ایست با استخراج

[1] Inclusion Index

[2] Proximity Index

[3] Equivalence

کلیدواژه‌های تک کلمه‌ای و چند کلمه‌ای، ماتریس هم‌رخدادی را تشکیل داده‌است. تحلیل‌های ایشان با استفاده از ابزار «رپیدماینر[۱]» و SNA[۲] بود است (Jalalimanesh ۲۰۱۲).

در سال ۲۰۱۲ «وی‌نینگ تو» به‌منظور کشف موضوعات نوظهور دو شاخص معرفی کرده است (شاخص‌های تازگی موضوع و تعداد انتشارات در آن موضوع). برای این منظور واژه‌هایی که را که بیشتر از ۳ بار در سال در یک کنفرانس یا مجله آمده‌اند انتخاب می‌شوند. یک موضوع کاندید، یک مجموعه‌ای از کلمات هستند که استخراج‌شده‌اند. یک موضوع کاندید نشان می‌دهد که یک واژه چقدر بااهمیت است. یک موضوع تحقیقاتی یک تقاطعی بین موضوع کاندید استخراج‌شده از مجله و کنفرانس است و درصورتی‌که در هر دو تکرار شود که موضوع تحقیقاتی محسوب می‌شود. یک موضوع تحقیقاتی داغ، موضوعی است که در چند باز زمانی متوالی مطرح‌شده باشد. هیچ‌وقت همه موضوع‌های تحقیقاتی داغ نهستند و در ضمن همیشه یک موضوع تحقیقاتی داغ نیست، و تنها درزمانی یک موضوع داغ محسوب می‌شود که در سطح بلوغ خود قرارگرفته باشد (ظهور، بلوغ، افول، و مرگ چرخه عمر فناوری هستند).

یک موضوع نوظهور یک موضوع تحقیقاتی است که مهم است ولی هنوز به موضوع داغ (بلوغ) تبدیل نشده است. تعریف دیگر PDY است که عبارت است از مدت‌زمانی که از اولین بار تا سال جاری است طول کشیده است که یک موضوع یک موضوع تحقیقاتی بشود و هیچ سالی دراین‌بین بدون مقاله نباشد در آن حوزه نباشد. شاخص نوآوری یا NI نیز برعکس PDY عمل می‌کند. برای مثال اگر موضوعی در ۵مین سال شروع خودش باشد شاخص نوآوری آن یک‌پنجم می‌شود. اگر موضوعی در بازه زمانی زیادی در ژورنال‌های مختلف موردبحث باشد آن موضوع بالغ شده است. PVI که شاخص تولید مقاله است، میزان تجمیع مقالات تا به سال مشخص‌شده به‌کل مقالات است. پیشنهادشده که نقطه تشخیص یا DP برای فناوری در حال ظهور نقطه‌ای باشد که دو نمودار به هم رسیده باشند و تازگی و میزان انتشار در نقطه‌ای برابر باشد (Tu and Seng ۲۰۱۲).

---

[۱] Rapid Miner

[۲] Social Network Analyser

در سال ۲۰۱۲ «گلنزل» تحلیل بر روی مقالات چاپ‌شده در سال‌های ۱۹۹۸ تا ۲۰۰۸ انجام داد که در آن در ابتدا با بررسی زمینه‌های تحقیقاتی مختلف، زمینه‌هایی که رشد زیادی داشته‌اند را انتخاب کرده و بعد از خوشه‌بندی این مقالات و تقسیم آن‌ها به زیرموضوع‌های تحقیقاتی، در طول بازه‌های زمانی ۵ ساله، برای کشف موضوع‌ها الگوهایی زیر موردبررسی قرار گرفت (Glänzel and Thijs۲۰۱۱):

- خوشه‌هایی که رشد زیادی دارند
- خوشه‌های جدیدی که تشکیل‌شده‌اند (این خوشه‌ها حتماً از ترکیب خوشه‌های دیگر به وجود آمده‌اند.
- خوشه‌هایی که موضوع آن‌ها عوض‌شده است

روش که برای کشف موضوع‌های نوظهور در سال ۲۰۱۴ توسط «اسمال» ارائه گردید، استفاده از خوشه‌بندی‌ها هم‌استنادی‌ها و استنادات مستقیم برای کشف این موضوع بوده است (Small, Boyack, and Klavans۲۰۱۴). منبع این تحقیق ۱۵ سال مقالات منتشرشده در اسکو پوس بوده است. با استفاده از استنادهای مستقیم کشف انجام خواهد شد و با استفاده از هم‌استنادی‌ها عملیات پالایه کردن انجام می‌شود. دو پارامتر زیر نیز برای اندازه‌گیری مورداستفاده قرارگرفته است:

- جدید بودن؛
- رشد زیاد.

در سال ۲۰۱۸ «استفان کارلی» ، روش و شاخصی برای تشخیص نوظهور بودن ارائه کرده است. در این روش که می‌توان هم بر روی مقالات و هم پتنت‌ها انجام شود، در مرحله اول، واژه‌ها استخراج می‌شوند و بعد از اعمال پالایش‌های فازی، تزاروسی و حذف کلمات اضافی، با استفاده از شاخص ارائه‌شده نوظهورها را تشخیص می‌دهد. شاخص ارائه‌شده در این طرح شاخص نوظهوری[۱] نام دارد. این شاخص از یک‌زمان مبدأ شروع می‌کند و با دنبال کردن زمان به دنبال الگوهای تازگی، رشد و پایداری می‌گردد. این شاخص با بررسی زمان ۷ ساله به پیدا کردن نوظهورها می‌پردازد. در این

[۱] Emergence Indicator (EI)

پژوهش بیان‌شده که شاخص نمی‌تواند در بدو تولید موضوعی به نوظهور بودن آن بپردازد البته افراد خبره این توانایی را دارند. برای ارزیابی خروجی‌ها به افراد خبره داده‌شده تا در خصوص منطقی بودن و آگاهی رساندن آن‌ها مورد تائید بشود. یکی از مشکلاتی که در این شاخص مطرح است استفاده نویسندگان مختلف از اصطلاحات متفاوت در یک حوزه است که در این صورت شاخص احتمالاً به‌خوبی عمل نکند (Carley et al. ۲۰۱۸).

در روش استفان کارلی شاخص EI در دو گام محاسبه می‌شود: پیدا کردن واژه‌های نوظهور، شمارش مواردی که (اشخاص، انتشارات، و ..) از آن واژه‌های نوظهور به‌صورت زیاد استفاده می‌کنند. برای محاسبه واژه‌های نوظهور به اختصار از روش زیر استفاده می‌شود:

- دوام و ماندگاری – واژه‌ای که در سه سال متوالی آمده باشد و حداقل هرسال در ۷ سند تکرار شده باشد.
- تازگی و رشد – واژه‌ها در زمان پایه کمتر از ۱۵٪ تکرار شده باشند، و در زمان‌های فعال بعدی بیش از دو برابر زمان قبل ظاهرشده باشند.
- اجتماع – واژه باید توسط دو فرد متفاوت نوشته‌شده باشند (این دو فرد نباید هم نویسنده باشند).
- رشد – از سه متغیر استفاده‌شده است
  - نسبت تغییرات سه سال آخر به سه سال ابتدایی دوره فعال
  - نسبت تغییرات دو سال آخر به دو سال قبل.
  - شب در وسط دوره فعال به آخرین سال (مثلاً سال ۷ تا سال ۱۰)

برای محاسبه‌ی اینکه چقدر یک بازیگر در تولید نوظهور نقش دارد به‌صورت زیر عمل خواهد شد:

- جزر ریشه دوم امتیازی که در مرحله قبل برای کلیدواژه نوظهور حساب شده را محاسبه کرده و در تعداد مقالاتی که آن کلیدواژه توسط همان بازیگر پیدا شده ضرب کنید، برای تمام کلیدواژه‌هایی که آن شخص استفاده کرده است این پارامتر را حساب کنید. سپس تمام امتیازات کلمات آن بازیگر را باهم جمع کنید.

- امتیاز نوظهوری مرحله قبل را با تقسیم بر جز ریشه دوم تعداد رکوردهای آن شخص نرمال کنید.

۲-۲-۵- موضوع‌بندی مقاله‌ها (مدل‌سازی موضوعی)

مدل‌سازی موضوعی یک دسته از روش‌های تحلیل متن است که جدیداً موردتوجه طیف وسیعی از محققین قرارگرفته است. ویژگی منحصربه‌فرد مدل‌سازی موضوعی این است که سعی در کد کردن محتوای یک مجموعه‌ای از متن‌ها به‌صورت مجموعه‌ای معنادار کدها که موضوع[۱] نامیده می‌شود را دارد. الگوریتم‌های این حوزه با حداقل دخالت انسان این امر را انجام می‌دهند و همین باعث جذاب شدن این موضوع نسبت به روش‌های تحلیل متن نسبت به روش‌های علوم انسانی شده است. در مدل‌سازی موضوعی بجای استفاده از عنوان‌های از قبل تعیین‌شده، با تعیین تعداد موضوع‌ها، موضوع‌ها از متون استخراج می‌شوند. همچنین مشخص می‌شود که هر سند با چه درصدی شامل کدام موضوع‌ها است (Mohr and Bogdanov۲۰۱۳).

حال سؤال اینجاست که چگونه می‌توان موضوع‌ها را از اسناد بیرون کشید و چه روش‌هایی برای این امر وجود دارد. پاسخ ساده این است که روش‌های استخراج موضوع براساس این فرض هستند که کلمات داخل متن از نظر معنایی وابسته به‌هم هستند. در این خصوص، معنایی که یک موضوع منسجم (در یک زمینه) را تشکیل می‌دهند از مجموعه کلمه‌های داخل خوشه تشکیل‌شده است. لذا در مدل‌سازی موضوعی در ابتدا صرف‌نظر از پیچیدگی‌های زبانی (مانند معانی یا مکان قرارگیری در متن) هم‌رخدادی کلمات را مورد تحلیل قرار می‌دهد. بنابراین هر سند به‌صورت مجموعه‌ای از کلمات مورد تحلیل و بررسی قرارگرفته می‌شود. هدف از مدل‌سازی موضوعی تحلیل این کلمات برای تشخیص الگوهای هم‌رخدادی کلمات در طول مجموعه عظیمی از دسته‌های کلمات (اسناد) و در انتها نگاشت توزیع کلمات در موضوع‌ها (و بالعکس) است. هم‌رخدادی کلمات می‌تواند ساختار فکری و زمینه تحقیقاتی یک حوزه را آشکار کند(Hu and Zhang۲۰۱۵).

[۱] Topic

استفاده از هم‌رخدادی کلمات در بسیاری از کارها مانند ترسیم ساختار حوزه‌های علمی (صدیقی ۱۳۹۳) بر این اساس

در سال ۲۰۰۸ «وارتنا» برای موضوع‌بندی اسناد از خوشه‌بندی کلیدواژه‌ها استفاده کرد. در پژوهش ایشان بعد از استخراج کلیدواژه‌ها با استفاده از خوشه‌بندی کا-میانه دوبخشی موضوع‌بندی را انجام شده است. برای ارزیابی نتایج را با دسته‌بندی موجود در ویکیپدیا مقایسه کرده و ارتباط مستقیم و قوی نشان می‌دهد.

در رشته‌های علم اطلاعات، مدل‌سازی موضوعی یک نمونه از مدل‌سازی آماری است. ساده‌ترین و پرکاربردترین مدل، LDA است که در سال ۲۰۰۳ توسط «بلیی و همکاران» معرفی‌شده است. این روش با این فرضیات کار می‌کند که هر سند به‌صورت مجموعه‌ای از کلمات (BOW) دیده می‌شود که ترکیبی از موضوع‌ها یا زمینه[1]ها را که در نظر نویسنده (برای بیان) بوده است را نشان می‌دهد. هر موضوع نیز توزیعی از کلمات پرتکرار (که نماینده هر سند هستند) است. بر اساس این دیدگاه نویسنده متن در برای بیان سند خود از موضوع‌هایی که می‌خواهد موردبررسی قرار دهد کلمات مرتبط را انتخاب کرده و در سند قرار می‌دهد تا موضوع‌های موردنظر خود را به‌صورت کامل بیان کند و سند به اتمام برسد. هدف از مدل‌سازی موضوعی پیدا کردن پارامترهای فرایند LDA است. بیشترین انتقادی که به روش‌های مدل‌سازی موضوعی گرفته می‌شود اتکای آن‌ها به بسته‌ای از کلمات است و ***ترتیب کلمات*** را در تحلیل‌های خود لحاظ نمی‌کنند (Blei, Ng, and Jordan ۲۰۰۳).

لازم به ذکر است که روش LDA روشی برای توصیف روابط بین اسناد و روش Vec۲Word روشی برای پیش‌بینی کلمات (به‌صورت محلی) است. با ترکیب این دو برداری جامع‌تر برای نمایش اسناد می‌توان نشان داد (Wang, Ma, and Zhang ۲۰۱۶). در اینجا می‌توان نتیجه گرفت که مدل‌سازی موضوعی متن می‌تواند از چند جنبه موردعلاقه باشد:

- این روش می‌تواند به محققین کمک کند تا محتوای اسناد بزرگ را به‌صورت خودکار کد کنند.

[1] Theme

- با این روش می‌توان پدیده‌های اجتماعی را مورد ارزیابی و اندازه‌گیری قرار دهیم.
- دید جدیدی برای پروژه‌های جدید می‌دهد.
- قابلیت بررسی متون کوچک (کمتر از نیم میلیون لغت) را فراهم می‌کند.

ازآنجاکه هر سند برخلاف برچسبی که دارد می‌تواند چندین عنوان موضوعی دیگر را داشته باشد (Schubert, Weiler, and Kriegel ۲۰۱۴)، پیدا کردن موضوع‌هایی برای یک سند بسیار موردتوجه است. هر سند می‌تواند با یک یا چند کلیدواژه خلاصه شود و محتوای اصلی آن سند به نمایش درآورده شود. هرچقدر تکرار کلمات بیشتر باشد اهمیت آن کلمه در آن سند بیشتر خواهد بود. هم‌رخدادی‌ها نیز می‌توانند ارتباط بین اسناد را نشان دهند (Dehdarirad, Villarroya, and Barrios ۲۰۱۴).

از طرفی به خاطر تنک بودن محتوای متن‌های کوچک چالش جدیدی در مدل‌سازی موضوعی اتفاق افتاده است. مدل‌های سنتی مانند LDA و pLSA فرض کرده‌اند که یکی سند بهارخوابی از موضوع‌های مختلف است که یک هر موضوع معنایی را بر اساس همبستگی بین کلمات ارائه کرده است و از آمار برای *به دست* آوردن موضوع بر اساس توزیع کلمات استفاده کرده‌اند. این روش‌ها بر روی‌داده‌های اسناد که هم‌رخدادی کلمات هستند کار می‌کند و درصورتی‌که سند کوتاه باشد ماتریس هم‌رخدادی کلمات تنک خواهد شد و تحلیل‌ها به‌صورت دقیق و کامل انجام نخواند گرفت (Cheng et al. ۲۰۱۴). راه‌حل‌هایی برای این مشکل **اسناد کوچک** ارائه‌شده است که برخی از آن‌ها به‌قرار زیر هستند:

- اسناد کوچک را با یکدیگر ترکیب کرده و یک شبه‌سند بزرگ *به دست* آوریم.
- اسنادی که کلیدواژه‌های مشابه دارند را ترکیب کرده و سند بزرگ‌تر *به دست* آوریم.
- با فرضیاتی از قبل تعیین‌شده متون کوچک را تحلیل کنیم.

«جیان هو» در سال ۲۰۰۸ با استفاده از روابط معنای ویکی-پدیا، خوشه‌بندی را بهبود داده است. در استخراج روابط از ویکی‌پدیا ضریب تأثیری از باارزش بودن اعمال کرده (مقالات سطح پایین‌تر

جزئیات بیشتری دارند از مقالات سطح بالاتر و ضریب بالاتری گرفته‌اند). اسنادی که ازنظر معنای شباهت باهم دارند در یک خوشه قرار دارند (Hu et al. ۲۰۰۸).

«یوته چانگ» در سال ۲۰۰۹ پژوهشی در خصوص تحلیل‌روند برای برچسب‌گذاری خوشه‌هایی که تشکیل داده است انجام داده است (در این الگوریتم اسناد خوشه‌بندی‌شده‌اند). برای این منظور بعد از حذف کلیدواژه‌های بدون بار معنایی در متن و استخراج کلمات مهم از عنوان و چکیده (که تکرار بیشینه‌ای دارند)، ضریب همبستگی کلمات را با هر خوشه محاسبه می‌کند و سپس بعد از مرتب کردن کلمات برحسب ضریب همبستگی، چند کلمه اول را به‌عنوان معرف خوشه (اسناد خوشه‌بندی شده‌اند) انتخاب می‌کند. برای خوشه‌های بزرگ از ضرب تکرار کلیدواژه در ضریب همبستگی استفاده‌شده است (Chang, Chang, and Tseng ۲۰۱۰). برای محاسبه ضریب همبستگی از رابطه زیر استفاده‌شده است:

$$Co(T,C) = \frac{TOP*TN - FN*FP}{\sqrt{(TP+FN)(FP+TN)(TP+FP)(FN+TN)}} \qquad (۵۵-۲)$$

«ژوئکی چنگ» در سال ۲۰۱۴ روشی برای مدل‌سازی موضوعی برای اسناد با متن کم ارائه کرده است. طرح ایشان بر این دو اصل استوار است که: از مدل‌سازی کلمات بر اساس هم‌رخدادی کلمات استفاده شود، و مقدار *به‌دست‌آمده* از هم‌رخدادی کلی کلمات در تمام متون در موضوع‌بندی اسناد استفاده شود (Cheng et al. ۲۰۱۴).

«نیلز نیومن» در سال ۲۰۱۴ با استفاده از دو روش به استخراج موضوع فنی و علمی تازه از مقالات علمی پرداخته است. در روش اول با استفاده از PCA و در روش دوم با استفاده از یک روش مدل‌سازی موضوعی این کار را انجام داده است. مراحل روش PCA به‌قرار زیر است: انتخاب رشته؛ پاک‌سازی اولیه (اسناد بی‌ربط را پاک می‌کند)؛ پاک‌سازی ثانویه (کلمات ایست، کلمات متداول علمی، کلمات به‌دردنخور، با استفاده از پایگاه داده‌های موجود)؛ پاک‌سازی اضافی؛ تثبیت و ترکیب (با استفاده از مفهوم term-clumping در ابتدا واژه‌های مرتبط با یکدیگر را پیدا می‌کند و در ادامه

با استفاده از الگوریتم‌های مبتنی بر قاعده[1] مترادف‌های کلمه را پیداکرده و آن‌ها را جایگزین می‌کنند؛ حرص کردن (کلمات توسط انسان بررسی می‌شوند و کلمات غیرمفید حذف می‌شوند)؛ تثبیت والد-فرزند؛ PCA. برای روش مدل‌سازی موضوعی از روش LDA استفاده کرده است.

در مقایسه بین این دو روش نشان داده است که روش PCA به‌صورت خیلی جزئی‌تر زیر-فناوری‌ها را مشخص و بر روی آن‌ها تمرکز کرده است. تحلیل LDA نیز دید متفاوتی از وضعیت فناوری به کاربر می‌دهد هرچند که کلید اصلی تحلیل برچسب‌گذاری هر موضوع است که امری انسانی است. درنهایت برای ارزیابی بهتر لازم و ضروری است که نظرات انسان هم در تحلیل‌ها لحاظ شود (Newman et al. ۲۰۱۴).

«شیرای بانرجی» در سال ۲۰۱۴ عملکرد الگوریتم کا-میانه و چند الگوریتم خوشه‌بندی مشتق شده از آن را باهم مقایسه کرده است (الگوریتم کامیانه دوبخشی، فازی، و ژنتیک). این مقایسه در داده‌های غیر متنی انجام‌شده است و عملکرد آن‌ها را در ۶ معیار اندازه‌گیری شده است. ایشان بیان کرده است که الگوریتم‌ها ازنظر کارایی به ترتیب الگوریتم کا-میانه ژنتیک، کا-میانه فازی، و کامیانه دوبخشی و سنتی بوده‌اند (Banerjee, Choudhary, and Pal ۲۰۱۵).

«ویوک کومار» در پژوهش خود در سال ۲۰۱۵ روشی برای مدل‌سازی موضوعی متن‌ها کوچک با استفاده از مدل مخلوطی گوسی[2] (برای بازنمایی کلمات) و خوشه‌بندی نرم ارائه کرده است. کومار بیان کرده است به خاطر تنک بودن داده‌ها روش‌های مانند LDA با مشکل مواجه می‌شوند، بنابراین این روش را ارائه کرده است. در این روش برای فیلترینگ کلمات از حداقل رخداد ۵ استفاده‌شده است که می‌تواند ضعف محسوب بشود (Sridhar ۲۰۱۵).

بنابراین یکی دیگر از ایراداتی از پژوهش کومار می‌توان گرفت یکپارچه‌سازی متن‌های کوچک برای تشکیل یک شبخ-متن بزرگ‌تر است که این کار به عمومیت ندارد (Quan et al. ۲۰۱۵). پژوهش (Quan et al. ۲۰۱۵)در سال ۲۰۱۵ نیز جهت رفع این مشکل ارائه‌شده است که با استفاده از متن‌ها کوتاه در یک حوزه پیشنهاد به رفع این مشکل داده است (با این منطق که کلمات

[1] Rule-based algorithm

[2] Gaussian mixture model

هم‌رخداد در یک حوزه بیشتر مورداستفاده است). برای ارزیابی روش ارائه‌شده با روش‌های پیشین از دو روش استفاده کرده است: استفاده از معیار انسجام[۱]، و استفاده از معیار خلوص[۲]. روابط زیر این نحوه محاسبه این دو معیار را نشان می‌دهد:

$$\mathrm{Coh_i} = \frac{۲}{k * \left(k - ۱\right)} \sum_{j<k\leq K} \log \frac{p(w_i, w_k)}{p(w_i) * p(w_j)} \qquad (۵۶-۲)$$

برای محاسبه انسجام مورداستفاده قرار می‌گیرد و K تعداد کلمات پرتکرار در هر خوشه است. $p(w_i, w_k)$ احتمال هم‌رخدادی کلمات w_i و w_j است. $p(w_j)$ و $p(w_i)$ احتمال رخداد دو کلمه در سندها هستند.

$$Purity = \frac{۱}{TK} \sum_{i} \max_{\mathrm{j}} \left|\tau_{z_i} \cap \tau_{g_j}\right| \qquad (۵۷-۲)$$

در رابطه بالا g_j و z_i به موضوع خاصی از متن‌ها کوتاه و بلند دلالت دارند و τ مجموعه K کلمه در موضوع‌های خاص هستند. در مقایسه‌های این مقاله بیان‌شده است که روش کا-میانه و LDA برای متن‌های کوتاه جواب قابل قبولی ارائه نمی‌کنند. یکی دیگر از راه‌های مقایسه این روش‌ها استفاده از دسته‌بندی است.

در پژوهش «یوسیم کیم» در ۲۰۱۷ برای تحلیل موضوعی[۳] که در (Kim et al. ۲۰۱۷) آمده است بیان شده که اگر تحلیل‌ها بدون استفاده از منابع زبان‌شناسی باشند نتایج *به‌دست‌آمده* احتمالاً غیرشفاف، با کلمات زیاد غیرکاربردی خواهند بود. «کیم» برای تحلیل موضوعی از ابر کلمات[۴] با برچسب تکرار هر کلمه استفاده کرده است. بعد از کشف موضوع می‌توان از آن در تحلیل‌روند، توصیف اسناد و هدایت اطلاعات[۵] استفاده کرد (Tu and Seng ۲۰۱۲) و (Jo, Lagoze, and Giles ۲۰۰۷).

---

[۱] Coherence

[۲] Purity

[۳] Topic Analysis

[۴] Word Cloud

[۵] Information navigation

۲-۲-۶- تحلیل‌روند

انجام هر نوع فعالیت علمی و عملی در یک حوزه خاص نیازمند آن است تا درک و فهم مناسبی از پژوهش‌های انجام‌شده در آن حوزه و روند تحقیقات آن حاصل شود. به‌طورمعمول سیاست‌گذاران، محققین و پژوهشگران در ابتدای شروع فعالیت خود در یک حوزه علمی، بامطالعه آثار پژوهشگران گذشته سعی بر درک ، روند حرکت و وضعیت جاری آن حوزه دارند. بر اساس میزان و عمق مطالعه، درک نسبی از چهارچوب کلی و وضعیت علمی جاری پیدا می‌شود. در راستای این درک، پژوهشگران از آخرین وضعیت پژوهشی علمی یک حوزه درک مناسبی پیداکرده و با در نظر گرفتن روند حرکت آن حوزه، موضوع‌های مناسب پژوهش برگزینند. سیاست‌گذار نیز با درک مناسب و دریافت نقشه علمی از وضعیت علمی و فناوری مربوطه می‌تواند با دیدی باز آینده علمی آن حوزه را ترسیم کرده و در خصوص آن سیاست‌های لازم را اتخاذ کرده و سرمایه گذاری‌های لازم را انجام دهند.

بنابراین تحلیل‌روند یک روش پیش‌بینی است که در آن با استفاده از برون‌یابی داده‌ها در طول زمان به وضعیت آینده را پیش‌بینی می‌کنند (Ziegler۲۰۰۹). لذا یک تحلیل‌روند خوب می‌تواند کمک شایانی برای مدیران و محققان انجام دهد و آن‌ها را در تصمیم‌گیری‌های آتی یاری رساند. ازاین‌رو تحلیل‌روند یک حوزه، برای پژوهشگران و سیاست‌گذاران آن حوزه لازم و ضروری است. تحلیل‌روند یک تحلیل فنی است که به‌منظور پیش‌بینی روند و حرکت چیزی مانند فناوری با استفاده از داده‌های گذشته انجام می‌شود (.Wu et al۲۰۱۱).

در تحلیل‌روند با استفاده از روش‌های آماری و ریاضی با استفاده از سری‌های زمانی داده‌ها را به آینده تعمیم می‌دهنده (.Wu et al۲۰۱۱). روش‌های تحلیل‌روند متفاوتی وجود دارد در همه آن‌ها فرض بر این است که داده‌ها به آینده نیز قابل تعمیم هستند (بخصوص در بازه‌های زمانی کوچک). نقطه ضعفی که این روش‌ها دارند نیاز به داده‌های زیاد برای کارایی مناسب است و بسیار در پیش‌بینی‌های دراز مدت آسیب‌پذیر هستند. از کاربردهای تحلیل‌روند می‌توان در تشخیص موضوع‌های نوظهور استفاده کرد (Wang۲۰۱۸). تحلیل‌روند کاربردهای زیر را می‌تواند داشته باشد (Rosenberg۲۰۰۷):

- تغییرات کلی الگو (ها) بر روی یک شاخص.
- مقایسه دو بازه زمانی متفاوت.
- مقایسه دو مکان مختلف در بازه زمانی.
- مقایسه دو جمعیت متفاوت.
- انطباق با آینده.

مستقل از هدف از تحلیل‌روند، قبل از شروع به تحلیل و تفسیر می‌بایست در تحلیل‌روند موراد زیر مشخص شوند (Kivikunnas۱۹۹۸):

- اندازه نمونه.
- وجود داده‌های پرت.
- وابستگی (همبستگی) شاخص‌های مورد ارزیابی با دیگر شاخص‌ها (بخصوص زمانی که اندازه داده‌ها کوچک باشند).

روش‌های مختلفی برای تحلیل‌روند یک حوزه علمی با استفاده از رویکرد آماری وجود دارد، طبق مطالعات انجام‌شده این روش‌ها را می‌توان به دو دسته کلی تحلیل بر روی اسناد علمی مانند (Kung et al. ۲۰۱۷, Müller et al. ۲۰۱۸) و تحلیل بر روی اختراعات انجام‌شده (Suh and Jeon ۲۰۱۷, Joung and Kim ۲۰۱۸) تقسیم کرد. ازآنجاکه در این رساله هدف تحلیل‌روند حوزه‌های علمی است، تحلیل اختراعات که در آن علم تبدیل به فناوری شده، موردبررسی قرار نخواهد گرفت. از طرفی هرکدام از این دسته‌ها می‌توانند بر اساس اطلاعات کتابشناختی و روش‌های متن‌کاوی مورد ارزیابی قرار گیرند. در تحلیل‌هایی که بر روی اختراعات ثبت‌شده انجام می‌شود علومی بررسی می‌شوند که به فناوری رسیده باشند. بنابراین در این نوع تحلیل از جامعیت تحلیل کاسته می‌شود. در تحلیل‌هایی که بر روی اسناد علمی انجام می‌شود تمام اسناد (چه علومی که به فناوری رسیده‌اند و چه علومی که هنوز به فناوری نرسیده‌اند) مرتبط با آن حوزه استخراج و موردبررسی قرار داده می‌شوند. در این نوع تحلیل، حجم زیاد اسناد پیچیدگی تحلیل را تا جایی بالا می‌برد که ممکن است باعث اشتباه در به دست آوردن نتایج شود. (Wu et al.۲۰۱۱).

تحلیل‌های کتابشناختی نوعی از تحلیل‌های آماری هستند که بر روی اسناد منتشرشده ازجمله مقاله و یا کتاب انجام می‌شود. عموماً این تحلیل‌ها بر روی اطلاعاتی که از این اسناد استخراج می‌شوند انجام می‌شود. این اطلاعات شامل عنوان، چکیده، نویسنده(ها)، زبان سند، نوع سند، کشور تولید کنند سند، سازمان تولیدکننده سند، استنادها و ... است. در این نوع تحلیل اطلاعات آماری از داده‌های موجود به‌دست‌آمده و تحلیل‌ها بر روی این اطلاعات انجام می‌شود. در تحلیل‌های متن‌کاوی، اطلاعات و الگوهای پنهان از روی داده‌های متنی بیرون کشیده می‌شود(Ozaydin et al.۲۰۱۷). در تحلیل‌روند حوزه علمی تحلیل متن‌کاوی می‌تواند بر روی عنوان، چکیده، استنادها و یا ترکیبی از آن‌ها انجام گیرد. در این پژوهش روشی ارائه می‌شود که با استفاده از تحلیل‌های کتابشناختی و متن‌کاوی، اطلاعات آماری استخراج‌شده را در چند مرحله پردازش و با استفاده از اطلاعات پردازش‌شده روند یک حوزه علمی را به‌طور خودکار به‌دست‌آمده و مورد تحلیل قرار گیرد.

همان‌طور که بیان شد تحلیل‌روند از اهمیت ویژه‌ای برخوردار است. برای این منظور روش‌هایی برای تحلیل‌روند ارائه‌شده است. روش‌های گوناگون تحلیل‌روند دارای نقاط قوت و ضعفی هستند که در بخش بعدی با جزئیات بیشتر به آن‌ها خواهیم پرداخت.

یکی از روش‌های اصلی تحلیل‌روند استفاده از ماتریس هم‌رخدادی است. روش‌های پیشین برای تشکیل ماتریس هم‌رخدادی از مجموع تعداد هم‌رخدادی کلیدواژه‌ها در اسناد مختلف به وجود آمده است (Viedma-Del-Jesus et al.۲۰۱۱) و (He۱۹۹۹). بنابراین از تکرار کلیدواژه‌ها که نشان‌دهنده میزان اهمیت کلمات است دخیل نشده است. این امر می‌تواند دقت ماتریس را کاهش دهد که در پی آن کاهش دقت در تشخیص روند و تحلیل‌های آتی است. از طرفی دیگر یکی از ابزارهای دیگر برای تحلیل‌روند، نمودار راهبردی است. برای تشکیل این نمودار نیز روش‌های مختلفی استفاده‌شده که در آن‌ها نیز تکرار کلیدواژه‌های تشکیل‌دهنده خوشه‌ها به‌صورت مستقیم دخیل نشده‌اند. استفاده از تکرار کلیدواژه‌ها می‌تواند در تحلیل‌ها اهمیت حوزه‌ها را لحاظ کند که در طرح پیشنهادی شاخصی بدین منظور ارائه گردید است.

یکی روش‌های تحلیل‌روند توسعه و تشکیل فیلدهای (پژوهشی) و تشخیص روابط بین آن‌ها استفاده از هم‌رخدادی کلمات است. روش‌های مختلفی برای تحلیل هم‌رخدادی کلمات وجود دارد که عبارت‌اند از: خوشه‌بندی کلمات هم‌رخداد، تحلیل شبکه‌های اجتماعی، و ***نمودار راهبردی*** (نمودار راهبردی می‌تواند پویایی تحقیق را نمایش دهد) اشاره کرد (Hu and Zhang ۲۰۱۵).

تشخیص روند[۱]، پیدا کردن تغییرات ناخواسته در برخی موارد مثل رخداد برخی موضوع‌ها در جریان داده‌ای است. اریک اسکابرد در (Schubert, Weiler, and Zimek ۲۰۱۵) به بررسی شباهت‌های تشخیص روند با استفاده از روش‌های تشخیص داده‌ها پرت پرداخته است. تشخیص روند کاربردهای گسترده‌ای در شبکه‌های اجتماعی ازجمله توییتر دارد. سیستمی که در (Mathioudakis and Koudas ۲۰۱۰) ارائه‌شده شامل سه مرحله است: گوش دادن به جریان داده‌ها و تشخیص کلیدواژه‌هایی که به‌صورت ناگهان رشد زیادی دارند، گروه کردن کلیدواژه‌ها، و درنهایت تحلیل‌روند این گروه‌ها.

روند یک الگو یا ساختاری در یک بعد است، البته روش‌هایی برای محاسبه چندبعدی آن نیز انجام‌شده است ولی این محاسبات غیرقابل‌دیدن است. روش‌های تحلیل‌روند به‌صورت زیر هستند (Kivikunnas ۱۹۹۸):

- روش‌های مبتنی بر تحلیل رگرسیون
- مدل زمانی پویای Warping
- تبدیل موجک
- تجزیه تحلیل زمانی کیفی شکل‌ها

همان‌طور که بیان شد مدل‌های موضوعی پیداشده می‌توانند در تحلیل‌روند مورداستفاده قرار گیرند. تحلیل‌روند با استفاده از موضوع‌های کشف‌شده شامل سه مرحله است (Tu and Seng ۲۰۱۲):

- تشخیص موضوع‌ها
- تشخیص موضوع‌های نوظهور و نحوه رشد آن‌ها

[۱] Trend Detection

- نمایش مشخصات موضوع

شاخص‌های مختلفی برای تحلیل هم‌رخدادی واژگان وجود دارد که می‌توانند مورد استفاده قرار گیرند. برای چگالی گره‌ها، تعداد روابط موجود و مرکزیت. مرکزیت درجه یکی از ساده‌ترین مرکزیت‌هاست که ارزش هر گره با شمارش تعداد همسایگانش به‌دست می‌آید. در شبکه هم‌رخدادی کلمات، هرچقدر مرگزیت درجه بیشتر باشد، ارتباط‌های بیشتری دارد و آن کلمه تاثیرگذارتر است. مرکزیت نزدیکی بر اساس مفهموم فاله و طول مسیر تعریف می‌شود. در شبکه، رئوسی که دارای حداقل فاصله با تمامی رئوس دیگر هستند، مرکزیت بالاتری دارند. مرکزیت بینیت بر اساس موقیت واژه‌ها در شبکه محاسبه می‌شود و کلمه‌ای که مقدار بیشتری داشته باشد راه‌های زیادی با تعداد بیشتری کلمه قرار دارد و ارتباط کلمات با استفاده از آن است. این کلمات یا موضوع‌ها قدرت ایزوله کردن یا افزایش ارتباط بین کلمات و موضوعات را بیشتری می‌کنند. موضوع‌ها و کلماتی که مرکزیت بینیت بیشتری دارند نقش حیاتی در ارتباط بین سایر گره‌ها را دارند و درصورتی که آن گره حذف شود جریان اطلاعات بین کلمات یا موضوع‌ها متوقف خواهد شد.(صدیقی ۱۳۹۳)

۲-۲-۶-۱- پیشینه پژوهش در تحلیل‌روند

همان‌طور که مطالعه تحقیقات گذشته نشان داد روش‌های گوناگونی بر روی حوزه‌های مختلف علمی برای تحلیل‌روند ارائه‌شده است. عموم این روش‌ها از هم‌رخدادی کلمات برای تحلیل‌روند و بررسی وضعیت استفاده کرده‌اند. در این روش‌ها با استخراج پارامترهای آماری ساده از روی فراداده‌های مقالات و تحلیل آن‌ها در بازه‌های زمانی متفاوت پرداخته‌اند. در برخی پژوهش‌های دیگر بعد از بررسی وضعیت آماری با استفاده از تحلیل هم‌رخدادی کلمات (رخداد دو کلمه در یک متن)، خوشه‌بندی و نمودار راهبردی (میزان ارتباط بین خوشه‌ها و کلمات داخل هر خوشه)، روند حوزه‌ها بررسی شده‌اند. به‌طورکلی از روش هم‌رخدادی کلمات برای تحلیل وضعیت، تحلیل‌روند و تحلیل‌های مقایسه‌ای مورداستفاده قرار می‌گیرند (Guo et al. ۲۰۱۷). (Callon et al. ۱۹۸۳).

بررسی پژوهش‌های گذشته نشان داد که در تشکیل ماتریس هم‌رخدادی تنها رخدادحضور (یا عدم حضور) دو کلمه در یک متن توجه شده است و میزان هم‌رخدادی کلمات در یک متن که نشان دهنده میزان اهیت یک کلمه است در نظر گرفته نشده است. به‌عنوان مثال اگر دو کلمه هر دو در یک متن یک بار تکرار شده باشند یا بیش از یک بار، میزان هم‌رخدادی آنها متفاوت نخواهد بود. در صورتیکه دو کلمه که در یک متن چندبار باهم آمده باشند، ارتباط نزدیک‌تری دارند نسبت به دو کلمه که در همان متن تنها یکبار تکرار شده‌اند. توجه به این گونه پارامترها می‌تواند میزان اهمیت کلیدواژه‌ها را در تحلیل‌های بعدی ازجمله خوشه‌بندی و بلوغ حوزه‌های علمی، بالاتر ببرد.

در پژوهش «ونگ» که در سال ۲۰۰۶ انجام شد با استفاده از روش مدل‌سازی موضوعی LDA استخراج موضوعی انجام شده است و موضوع‌های استخراج شده در یک توزیع زمانی مورد بررسی قرار می‌دهد. در صورتی که کلمات در یک بازه زمانی کوتاه به صورت قوی آمده باشند و دیگر نیامده باشند، موضوع یک توزیع ضعیف تشکیل می‌دهد.

فعالیت‌های بسیاری برای تحلیل‌روند تحقیقات در یک حوزه انجام‌شده است. در سال ۲۰۰۷ «چااو» روند حرکت فناوری RFID را در (Chao, Yang, and Jen ۲۰۰۷) در بازه زمانی ۱۵ سال بر روی مقاله‌های منتشرشده مجلات SCI[۱] انجام داده است. در پژوهش ایشان، اسنادی که در این مقالات در عنوان و یا چکیده خود، واژه RFID و یا Radio Frequency identification را بکار برده‌اند از پایگاه داده WOS استخراج‌شده است. این پژوهش تنها به تحلیل اطلاعات آماری بسنده کرده و روند حرکتی یک حوزه علمی را بررسی نکرده است. بنابراین امکان توانایی در تشخیص میزان بلوغ رشته‌ها و حوزه‌های تازه را ندارد. تحلیل‌های انجام‌شده عبارت‌اند از: میزان همکاری بین‌المللی، نویسندگی، زبانی، نوع سندی، و تعداد استنادها.

«چوی» در سال ۲۰۰۹ در (Choi and Park ۲۰۰۹) مسیر حرکت فنّاوری ساختارهای ارگانیک[۲] را بررسی کرده است. این تحلیل بر اساس هم‌استنادی اختراعات انجام‌شده است. بر این اساس اسناد با استفاده از روش خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی شده‌اند. البته در این تحلیل از کلیدواژه‌ها

[۱] Science Citation Index

[۲] Organic

که در حقیقت ویژگی اصلی تحلیل‌روند یک حوزه علمی به شمار می‌رود، استفاده نشده است. به خاطر عدم بررسی کلیدواژه‌های مقاله‌ها مفاهیم و محتوای پنهان درون مقاله‌ها بررسی نشده است و توانایی کشف میزان بلوغ رشته‌ها به‌صورت دقیق ندارد.

«چنگ» در سال ۲۰۱۰ روند را در حوزه علم تحصیل به‌صورت خودکار موردبررسی قرار داده است (روش جدید ارائه نکرده تنها موردبررسی قرار داده). این تحلیل شامل شش گام است؛ بخش‌بندی متن، پیدا کردن شباهت بین متون، خوشه‌بندی چند سطحی، برچسب‌گذاری خوشه‌ها، تحلیل Facet، بصری سازی. در مرحله خوشه‌بندی بعد از حذف داده‌های پرت، از خوشه‌بندی سلسله‌مراتبی استفاده‌شده است. برای برچسب‌گذاری خوشه‌ها بعد از حذف کلیدواژه‌های بدون بار معنایی در متن و استخراج کلمات مهم از عنوان و چکیده، ضریب همبستگی کلمات را با هر خوشه محاسبه می‌کند و سپس بعد از مرتب کردن کلمات برحسب ضریب همبستگی، چند کلمه اول را به‌عنوان معرف خوشه انتخاب می‌کند. برای خوشه‌های بزرگ از ضرب تکرار کلیدواژه در ضریب همبستگی استفاده‌شده است (Chang, Chang, and Tseng ۲۰۱۰).

«صدیقی» در سال ۲۰۱۱ بعد از استخراج مقالات حوزه مدیریت دانش از پایگاه «وب آوساینس» و با استفاده از از نرم افزارهای SPSS و «هیست سایت» به مطالعه روند آن حوزه در بین سال‌های ۲۰۰۱ تا ۲۰۱۰ و ترسیم ساختار آن حوزه پرداخته است. ایشان اسناد را در خوشه‌هایی دسته‌بندی کرده و چهار خوشه اصلی را جدا کرده و ارتباط بین آن‌ها را مورد تحلیل قرار داده است. نقشه ترسیم علوم‌نگاشتی بر اساس استنادهای »جی سی اس» برای هر خوشه انجام شده است (صدیقی and جلالی‌منش ۱۳۹۱).

«پارک» در سال ۲۰۱۰ در (No and Park ۲۰۱۰) نفوذ و روند فناوری نانو را موردبررسی قرار داده است. پژوهش وی با استفاده از تحلیل هم‌استنادی بر روی اختراعات انجام‌شده است. در این پژوهش شاخصی تعریف‌شده که بر اساس آن با استفاده از استنادهای مستقیم میزان نفوذ فناوری‌ها روی یکدیگر نشان داده‌شده‌اند. این تحلیل از سه مرحله جمع‌آوری داده، تشکیل ماتریس هم‌رخدادی استنادی و دسته‌بندی آن‌ها، و تحلیل نفوذ تشکیل‌شده است. در پژوهش وی، از نمایه‌هایی که برای هر اختراع ثبت‌شده استفاده‌شده و محتوای متن اختراع و کلمات مهمی که در

آن نویسنده آن اختراع استفاده کرده، در نظر گرفته نشده است. از طرفی دیگر میزان دفعات رخداد کلیدواژه‌ها در محاسبه هم‌رخدادی لحاظ نشده است. تحلیل هم‌استنادی به خاطر عدم توجه به محتوای متن مقاله باعث چشم‌پوشی از برخی فناوری‌های نوظهور خواهند شد.

در سال ۲۰۱۱، «پنگ» در (Lv et al. ۲۰۱۱) تحلیل کتابشناختی را بر روی مقالاتی که در ۵ سال متوالی بر حوزه گرافن[۱] در WOS به چاپ رسیده انجام داده است. تحلیل انجام‌شده در دو مرحله استخراج اطلاعات از منابع داده و تحلیل‌های کتابشناختی انجام‌گرفته است. تحلیل‌های انجام‌شده با استفاده از نرم‌افزار TDA[۲] و تنها بر روی اطلاعات کتابشناختی بوده است. بنابراین اطلاعاتی که توسط نویسنده در چکیده و عنوان مستتر است، نادیده گرفته‌شده است. عدم استفاده از متن چکیده و عنوان باعث کاهش دقت در به دست آوردن میزان بلوغ رشته‌ها و حوزه‌هایی است که در متن به‌صورت ضمنی به آن‌ها اشاره شده است.

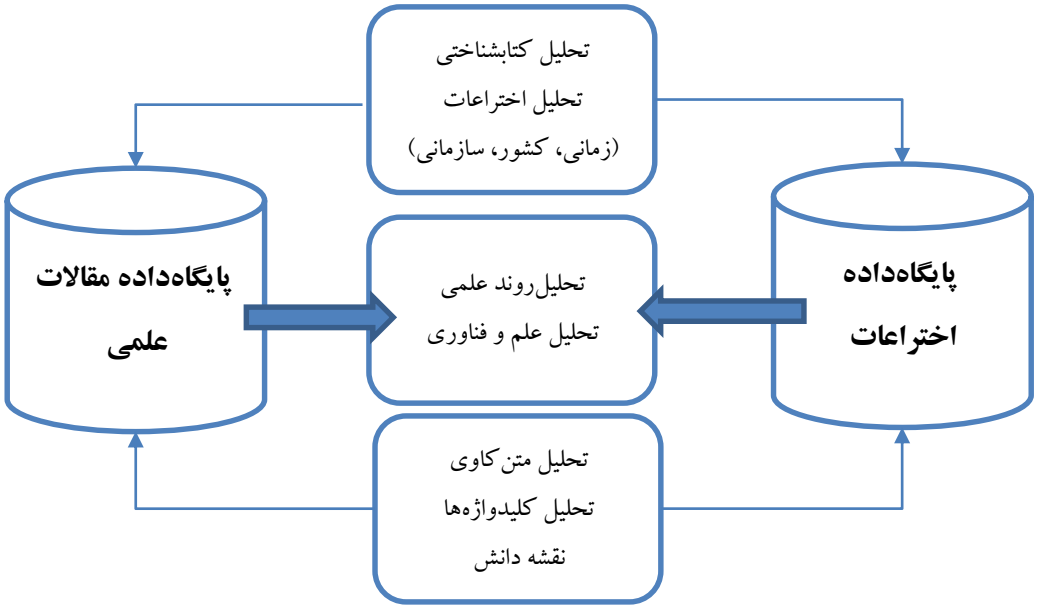
در سال، ۲۰۱۱ «فنگ» در (Wu et al. ۲۰۱۱) برای تحلیل‌روند «اتچینگ[۳]» یک روش سیستمی ارائه کرد و برای تحلیل‌روند از سه روش تحلیل کتابشناختی، تحلیل اختراع، و متن‌کاوی استفاده نمود. برای تحلیل اختراع از نرم‌افزار Delphion و برای متن‌کاوی از نرم‌افزار Tseng بهره گرفته‌شده است. در این پژوهش تنها از ۵۰ کلمه پرتکرار اول استفاده‌شده و مابقی کلمات نادیده گرفته‌شده‌اند. فراوانی کلمات هر خوشه با یکدیگر جمع شده و در بازه‌های زمانی یکسانی فراوانی‌های کل خوشه‌ها نمایش داده‌شده است. ازآنجاکه تحلیل متن‌کاوی هم بر روی عنوان و هم بر روی چکیده انجام‌شده، این امکان وجود دارد که کلماتی که در نوآوری مقاله نقشی ندارند و مربوط به فناوری‌های گذشته هستند و برای تعیین ارتباط فناوری‌های پیشین با یک پژوهش در متن چکیده آمده‌اند، در تحلیل‌ها اثر زیادی بگذارند و نتایج تحلیل را منحرف سازند. بنابراین استفاده از ۵۰ کلیدواژه نخست و استخراج و استفاده از کلمات پرتکرار حوزه‌های پایه باعث و نادیده دیگر کلیدواژه‌ها باعث کاهش دقت در تحلیل‌روند و عدم کشف حوزه‌های جدید خواهد شد.

---

[۱] Graphene

[۲] Thomson Data Analyzer

[۳] Etching



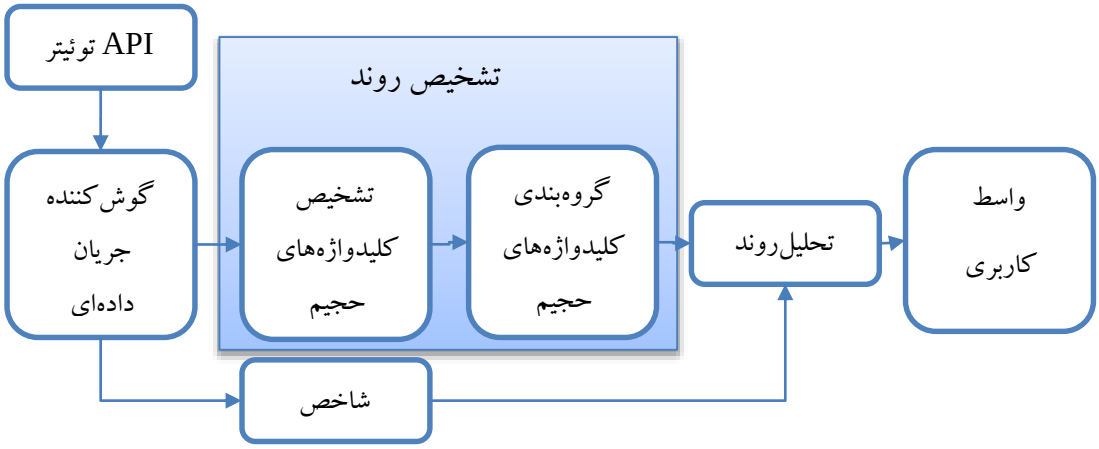
شکل ۲-۱۳ ساختار پژوهش (Wu et al. ۲۰۱۱)

تحلیل دیگری برای روند سلول‌های بنیادی در سال ۲۰۱۱ توسط «زین» با استفاده از هم‌رخدادی کلمات و وزن دهی به عنوان‌های موضوعی[۱] در (An and Wu ۲۰۱۱) انجام‌شده است. در این تحلیل که بر مبنای روش نیمه‌خودکار است با استفاده از آنتروپی اطلاعات کلمات بااهمیت استخراج‌شده و با استفاده ازنظر متخصصان حوزه برای استخراج عناوین موضوعی استفاده‌شده است. تحلیل انجام‌شده با استفاده از نمودار راهبردی صورت گرفته است. از طرفی دیگر میزان دفعات رخداد کلیدواژه‌ها در محاسبه هم‌رخدادی لحاظ نشده است و تنها از دو وزن ۱ و ۲ برای عنوان‌های موضوعی استفاده‌شده است. استفاده از متخصصان در انتخاب عناوین موضوعی باعث کندی در استخراج عناوین و جهت گیری این عنوان‌ها در تحلیل‌ها می‌شود.

$$Growth\ Rate = \frac{Current\ Year\ Total - Previous\ Year\ Total}{Previous\ Year\ Total} \qquad (۵۸-۲)$$

[۱] Subject heading

در سال ۲۰۱۲ نیز برای اولین بار «مایکل متیوداکیس» روشی برای تشخیص روند بر روی جریان داده‌های توییتر ارائه کرد (Mathioudakis and Koudas ۲۰۱۰). این روش شامل سه مرحله بوده است: تخیص کلیدواژه‌های یهویی (که به‌صورت ناگهانی با نرخ زیادی در توئیت‌ها ظاهر می‌شوند، گروه‌بندی کلیدواژه‌های یهویی با دیگر کلیدواژه‌ها بر اساس هم‌رخدادی، و به دست آوردن اطلاعات اضافی با استفاده از توئیت‌هایی که متعلق به توئیت‌های کلیدواژه‌های ترند بوده‌اند.



شکل ۲-۱٤ تحلیل روند در توئیتر(Mathioudakis and Koudas ۲۰۱۰)

«ایوان مارچین» در سال ۲۰۱۲ روشی برای تعیین روند موضوعات ارائه کرده است. برای این منظور در ابتدا، از توزیع کلمات برای بازنمایی توئیت‌ها در یک بازه زمانی خاص استفاده کرده و رابطه معنایی بین توئیت‌ها و موضوع‌های ویکی‌پدیا را به دست آورده است. سپس، با استفاده از همبستگی موضوعی[۱] ارتباط بین موضوع‌ها را بر روی زمان به دست آورده است. در انتها ارتباط بین موضوع‌های ترند با موضوع‌های فعالیت‌های توئیت به دست آورده است. برای ارزیابی کار نیز از میزان جستجوی موضوع‌های مختلف از گوگل‌ترند استفاده کرده است (Marcin and Shiu ۲۰۱۲).

در سال ۲۰۱۳ «هو چنگ» در پژوهش خود به تحلیل مقالات علم اطلاعات پرداخته است. با استفاده از روش خوشه‌بندی سلسله‌مراتبی و با فرض اینکه هر خوشه یک زیرشاخه‌ای از آن حوزه

[۱] Topic Correlation

است، با استفاده از نمودار راهبردی روند حوزه علم اطلاعات در طول ۵ سال موردبررسی قرارگرفته است (.Hu et al ۲۰۱۳). در تحلیل درون خوشه‌ای راهبردی مورداستفاده تعداد ارتباطات بین کلیدواژه‌ها شمارش‌شده که ارزش و وزن خود کلمه صرف‌نظر شده است. در همان سال «سمند کوچکسرایی» روند تحقیقات سلول‌های بنیادی در کشور ایران و قاره‌های اروپا و امریکا و در منطقه خاورمیانه را تنها با استفاده از اطلاعات آماری مقالات چاپ‌شده در PubMED موردبررسی قرار داده است (Samadikuchaksaraei, Mohammadhassanzadeh, and Shokraneh ۲۰۱۳). این تحلیل روشی برای *به دست* آوردن زیر حوزه‌ها و سطح بلوغ رشته‌ها روشی ارائه نکرده است.

علاوه بر تحقیقات فوق، با استفاده از تحلیل هم‌رخدادی و نمودار راهبردی در سال ۲۰۱۵ تحلیل‌روند در خصوص میزان تمرکز، بلوغ و توسعه تحقیقات بر روی «سامانه‌های توصیه‌کننده» انجام‌شده است (Hu and Zhang ۲۰۱۵). استفاده تعداد لینک‌ها در نمودار راهبردی باعث شده است که میزان تأثیر وزن کلمات در خوشه در نظر گرفته نشود و این امر باعث صرف‌نظر کردن اهمیت کلیدواژه‌ها و کاهش دقت تحلیل خواهد شد.

«چن» نیز در سال ۲۰۱۶ با استفاده از نرم‌افزار VOSviewer تحلیل هم‌رخدادی کلمات و مطالعه روابط بین موضوع‌های پروژه‌های مهندسی و علوم مدیریت که در دوره ۵ ساله در کشور چین انجام‌داده است (.Chen et al ۲۰۱۶). روشی برای انتخاب تعداد کلیدواژه‌های مورد تحلیل و روابط بین خوشه‌ها ارائه نشده و مورد تحلیل قرار نگرفته است. از طرفی دیگر میزان دفعات رخداد کلیدواژه‌ها در محاسبه هم‌رخدادی لحاظ نشده است. که لحاظ کردن این تکرار می‌تواند میزان اهمیت و وابستگی دو کلیدواژه را نسبت به یکدیگر نشان دهد.

«چنگ» در ۲۰۱۷ برای پیدا کردن موضوع‌های تحقیق‌های موردتوجه دانشجویان دندان پزشکی از تحلیل هم‌رخدادی کلمات استفاده کرده است. وی این کار را با استفاده از نرم افزار «بیب اکسل»[۱] و استخراج کلیدواژه‌های با تکرار بالاتر و تحلیل خوشه‌بندی انجام داده است. کلیدواژه‌های پرتکرار از رابطه T ارائه‌شده در مقاله استخراج شدند. در انتها با ترسیم تنها شکلی از خوشه‌بندی

[۱] BIbExcel

کلیدواژه‌های پر تکرار موضوع‌های مورد علاقه را نشان داده است (Chang et al. ۲۰۱۷). در رابطه زیر I_1 تعداد واژه‌هایی با تکرار ۱ است.

$$T = \frac{-1 + \sqrt{1 + 8 * I_1}}{2} \qquad (۵۹-۲)$$

«ژائو فنگون» نیز برای تحلیل‌روند از نمودار راهبردی و شبکه‌های اجتماعی استفاده کرده است. در روش پیشنهادی فنگون از کلیدواژه‌هایی که بیش از ۲۰ بار تکرار شده‌اند استفاده‌شده است. این امر روش انتخاب می‌تواند به نادیده گرفتن کلیدواژه‌های مهم و یا وارد کردن کلیدواژه‌هایی غیر مرتبط شود (Zhao et al. ۲۰۱۸)

همان‌طور که مطالعه تحقیقات گذشته نشان داد روش‌های گوناگونی بر روی حوزه‌های مختلف علمی برای تحلیل‌روند ارائه‌شده است. عموم این روش‌ها از هم‌رخدادی کلمات برای تحلیل‌روند و بررسی وضعیت استفاده کرده‌اند. در این روش‌ها با استخراج پارامترهای آماری ساده از روی فراداده‌های مقالات و تحلیل آن‌ها در بازه‌های زمانی متفاوت پرداخته‌اند. در برخی پژوهش‌های دیگر بعد از بررسی وضعیت آماری با استفاده از تحلیل هم‌رخدادی کلمات (رخداد دو کلمه در یک متن)، خوشه‌بندی و نمودار راهبردی (میزان ارتباط بین خوشه‌ها و کلمات داخل هر خوشه)، روند حوزه‌ها بررسی شده‌اند. به‌طورکلی از روش هم‌رخدادی کلمات برای تحلیل وضعیت، تحلیل‌روند و تحلیل‌های مقایسه‌ای مورداستفاده قرار می‌گیرند (Guo et al. ۲۰۱۷). (Callon et al. ۱۹۸۳).

بررسی پژوهش‌های گذشته نشان داد که در تشکیل ماتریس هم‌رخدادی تنها رخدادحضور (یا عدم حضور) دو کلمه در یک متن توجه شده است و میزان هم‌رخدادی کلمات در یک متن که نشان دهنده میزان اهیت یک کلمه است در نظر گرفته نشده است. به‌عنوان مثال اگر دو کلمه هر دو در یک متن یک بار تکرار شده باشند یا بیش از یک بار، میزان هم‌رخدادی آنها متفاوت نخواهد بود. در صورتیکه دو کلمه که در یک متن چندبار باهم آمده باشند، ارتباط نزدیک‌تری دارند نسبت

به دو کلمه که در همان متن تنها یکبار تکرار شده‌اند. توجه به این گونه پارامترها می‌تواند میزان اهمیت کلیدواژه‌ها را در تحلیل‌های بعدی ازجمله خوشه‌بندی و بلوغ حوزه‌های علمی، بالاتر ببرد.

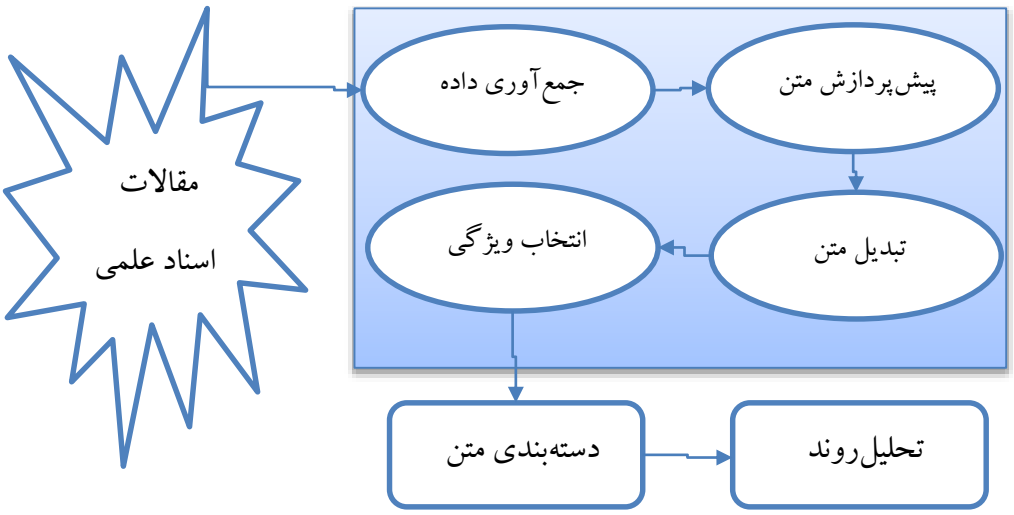
از طرفی دیگر در تحلیل‌های انجام‌شده در نمودار راهبردی به تعداد لینک‌های ارتباطی بین خوشه‌ها و کلمات داخل خوشه توجه شده است و فراوانی تکرار کلمات هر خوشه در میزان چگالی خوشه موردتوجه قرار نگرفته است. در بخش بعدی به بیان روشی خواهیم پرداخت که می‌تواند روش‌های موجود در تحقیقات گذشته را با در نظر گرفتن شاخص‌های جدیدی بهبود دهد.

در سال ۲۰۰۱ «رجارامان» روشی براش تشخیص و دنبال کردن موضوع‌ها و تحلیل‌روند ارائه کرد. در این روش بعد از دریافت جریان متن و اعمال یکسری پیش‌پردازش‌ها (مانند حذف کلیدواژه‌هایی که کمتر از ۵٪ کل متن را شامل بشوند و انتخاب n کلیدواژه برتر)، با استفاده شبکه عصبی موضوع‌ها را کشف کرده و بعد از حذف خوشه‌های (موضوع‌ها) مشکوک (که احتمال آن‌ها کم باشد) هر سند را به هر موضوع (یا چند موضوع) نسبت می‌دهد (Rajaraman and Tan ۲۰۰۱).

از طرفی دیگر ازآنجاکه در یک سند یک عنوان اصلی و چند عنوان فرعی می‌تواند وجود داشته باشد که عنوان اصلی موردتوجه و ترند نباشد اما موضوع فرعی روندی فعال و موردتوجه داشته باشد (Schubert, Weiler, and Kriegel ۲۰۱۴). «شوبرگ» در (Schubert, Weiler, and Kriegel ۲۰۱۴) در سال ۲۰۱۴ موضوع‌های ترند را زیرموضوع‌هایی بیان می‌کند که به‌صورت خوشه‌های با اندازه کوچک هستند و رشد قابل توجه‌ی دارند. البته ازنظر ایشان موضوع جدید، موضوعی است که از کلمات *محبوب (تکرار بالا)* و *جدید* استفاده کرده است. برای کشف موضوع‌های ترند از روش آماری z-score استفاده کرده است. در مقاله اشاره‌شده نیز بیان‌شده که موضوع‌های در حال ظهور از داده‌های پرت تشکیل‌شده است و برای محاسبه آن‌ها با گرفتن میانگین و واریانس داده‌ها (تشکیل مدل مرجع) داده‌های پرت را می‌تواند تشخیص دهد. (بنابراین روش‌ها سلسله مراتبی می‌توانند به‌صورت دقیق‌تری زیر خوشه‌ها را تحلیل و کشف کنند.).

ازآنجاکه چکیده داده‌های لازم را برای تحلیل فراهم می‌کند، تحلیل کلیدواژه‌ها و چکیده برای روند کافی است و داده‌های دیگری که در متن مقاله‌ها، جداول و شکل‌ها وجود دارد لزومی ندارد

(۲۰۱۵Ojo and Adeyemo). «آدبولا» در ۲۰۱۵ قالبی که برای کشف ترند مورداستفاده قرارگرفته است به‌صورت زیر است.



شکل ۲-۱۵ قالب کاری کشف روند (۲۰۱۵Ojo and Adeyemo)

برای محاسبه موضوع‌های جدید (۲۰۱۲Lau, Collier, and Baldwin) در ابتدا با استفاده از LDA موضوع‌ها و توزیع کلمات را بر روی آن موضوع‌ها در بازه‌های زمانی متفاوت محاسبه کرده است. در ادامه درجه تغییرات (یا تکامل یک موضوع) را با استفاده از معیار واگرایی «جنسن–شانون»[۱] بین توزیع کلمات هر موضوع (مثلاً t) را قبل و بعد از بروز شدن (توزیع کلمات) محاسبه کرده است. درصورتی‌که این پارامتر از حد آستانه‌ای بیشتر بشود آن موضوع جدید محسوب می‌شود. پژوهش مذکور بر روی داده‌های توییتر انجام‌شده است.

پژوهش دیگری برای تحلیل‌روند بر روی متن‌های مقالات بیومدیکال با استفاده از استخراج هم‌معنی‌ها توسط «بوسکانیو» در سال ۲۰۱۴ انجام شد. در این پژوهش تمرکز بر روی پیداکردن ترندهای زمانی در این حوزه شده است. این ترندها با استفاده از تغییر کلمات در گذر زمان شناسایی می‌شوند. درصورتی‌که بردار کلمه (که ارتباط با دیگر کلمه‌ها را نشان می‌دهد) در طول زمان تغییرات زیادی بکند به‌عنوان ترند زمانی درنظر گرفته می‌شود. بنابراین باید غیر از بردار فعلی کلمه بردارهای قبلی و بعدی نیز محاسبه شوند (۲۰۱۴Buscaino).

[۱] Jensen-Shannon divergence

برای محاسبه ترندهای زمانی در (Buscaino۲۰۱۴) دو راه پیشنهادشده است: استفاده از روش آشفتگی[۱] و ماتریس تبدیل[۲]. در روش آشفتگی در هنگام محاسبه بردار معنایی کلمه در زمانی مشخص بردار را ذخیره کرده و سپس به ادامه آموزش بردار کلمه می‌پردازند. بعد از آموزش بردار جدید را با بردار قدیم مقایسه می‌کنند تا میزان تغییر بردار بررسی شود.

«کیم» در سال ۲۰۱۷ راهکاری چهار مرحله‌ای با استفاده از آنتولوژی برای تحلیل‌روند علمی و فناوری ارائه کرده که این مراحل عبارت‌اند از؛ جمع‌آوری داده، پردازش زبان طبیعی، تحلیل متن، و نمایش. در پردازش زبان طبیعی از روش ۵ مرحله‌ای برای تولید آنتولوژی استفاده کرده است که در آن با استفاده از متخصصین این آنتولوژی ساخته می‌شود. با توجه به اینکه تحلیل سری زمانی یکی از تحلیل‌های روند پرطرفدار برای تحلیل موضوع‌های در حال ظهور محسوب می‌شود، «کیم» برای تحلیل‌روند با استفاده از سری زمانی این تحلیل را انجام داده است (Kim et al. ۲۰۱۷). نمودارهای ارائه‌شده در تحلیل کیم با استفاده از تکرار کلمات نشان داده‌شده‌اند. کیم روش‌شناسی چهار مرحله‌ای خود را برای تحلیل‌روند، با استفاده از یک نمونه اعتبارسنجی[۳] کرده است.

اگر در تحلیل شما روندی را شما پیدا کردید که به ناگهان سقوط کرده است این امکان دارد که شخصی باسابقه بالا آن رشته را ترک کرده و جایگزینی نیز ندارد (van Raan ۱۹۹۹).

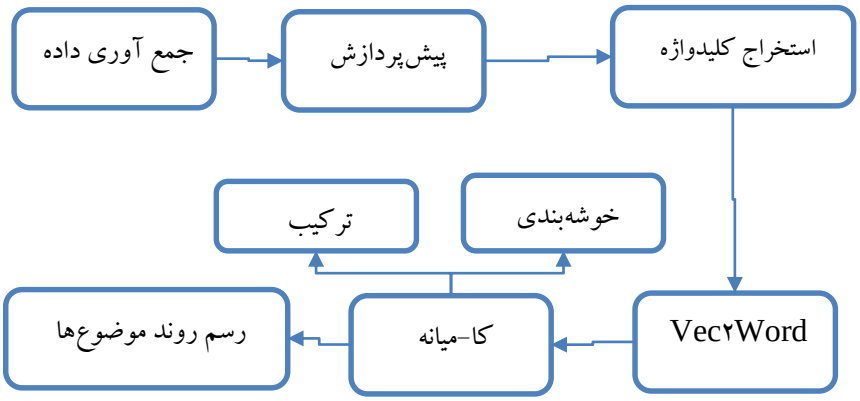
پژوهش سال ۲۰۱۷ (Tao et al. ۲۰۱۷) توسط «هونگ‌ژی» برای پیدا کردن موضوع‌های جدید روشی پیشنهاد داده است که با استفاده از کا-میانه وزن‌دار این تحلیل‌روند را انجام می‌دهد. در این روش از ترکیب vec۲word و استخراج کلیدواژه استفاده کرده است. بعد از استخراج کلیدواژه‌ها، کلیدواژه‌های مرجع (به‌عنوان موضوع اصلی) را پیدا می‌کند و با استفاده از فاصله کوسین، کلیدواژه‌های فرعی را بر اساس کلیدواژه‌های اصلی بیان می‌کند.

---

[۱] Perturbation Method

[۲] Transformation Matrix

[۳] Validity



شکل ۲-۱۶ مدل ساختاری ارائه‌شده در (.Tao et al ۲۰۱۷) برای کشف موضوع‌های جدید

ازآنجاکه داده‌ها مربوط به داده‌های واقعی هستند و هیچ معیاری برای تشخیص کارایی روش وجود ندارد از عناوین WHO (روز جهانی بدون تنباکو) استفاده کرده است. در آزمایش انجام‌شده برای هر سند ۱۰ واژه استخراج کرده است و برای هر واژه بردار ۱۰۰عنصری و به خاطر اینکه در کنفرانس WHO ۵ موضوع بررسی‌شده است تعداد خوشه‌ها را نیز ۵ در نظر گرفته است. از معایب این روش انتخاب کلمه مرجع به‌صورت دستی و انتخاب تعداد خوشه‌ها به‌صورت تصادفی و بدون حساب است.

«سهیلی» در سال ۱۳۹۷ در پژوهش خود به شناسایی و مقایسه روند موضوعی مفاهیم حوزه علم اطلاعات و دانش‌شناسی ایران در دو بازه ۵ ساله پرداخته است. ایشان با استفاده از روش‌تحلیل هم‌رخدادی واژگان، همه مجله‌های فارسی زبان در حوزه اشارهشده را که در مجلات معتبر به چاپ رسیده‌اند را مورد تحلیل قرار داده است (سهیلی, خاصه, and کرانیان ۱۳۹۷).

«شناهو ژیانگ» در پژوهش خود در ۲۰۱۸ با استفاده از این اصل که عبارات مهم در یک زمینه توسط افراد مهم آن زمینه نوشته می‌شوند، یک چهارچوبی برای تشخیص روند ارائه کرده است. این کار توسط امتیازدهی به عبارات مهم و نویسندگان مهم انجام‌شده است. سپس برای محاسبه امتیاز زمینه‌های تحقیقاتی‌ترند، بعد از امتیازدهی این عبارات را به یک شبکه عصبی بازخوردی برای داده است. بر این اساس برای ارزیابی کار خود روندهای گذشته را مورد ارزیابی و پیش‌بینی قرار داده است. به این صورت که با عوض کردن روش امتیازدهی به کلمه‌های کلیدی با روش

TextRank پیش‌بینی‌ها را باهم مقایسه کرده است!!!. بنابراین تشخیص روند به در پژوهش شناهو شامل سه مرحله زیر است (Marcin and Shiu ۲۰۱۲, Jiang et al. ۲۰۱۸):

- استخراج کلیدواژه‌های کلیدی
- بهبود در بازنمایی با استفاده از بازنمایی کلمات (بازنمایی کلمات) (شاخص tf*iaf که بر اساس نویسنده بیان‌شده است)
- پیش‌بینی و امتیازدهی به عبارات با استفاده از RNN

مشکلی که در این روش می‌تواند به وجود بیاید این است که اشباع شدن یک‌رشته در نظر گرفته نشده است و در ضمن امتیازدهی به نویسندگان برتر می‌تواند بر اساس بایاس فکری باشد و در ضمن افرادی مثل «میکولو» که معروف نیستند با ابداع روشی ترند را تغییر بدهند.

در پژوهش (Wu et al. ۲۰۱۸) در سال ۲۰۱۸ برای فراهم کردن یک نقشه از تحقیقات بر روی چند سرطان کشنده از LDA استفاده کرده است. به‌این‌ترتیب که ابتدا مقالات منتشرشده در LiveScience را استخراج کرده است و سپس عملیات پیش‌پردازشی برای استخراج عنوان و چکیده هر مقاله انجام داده است. در ادامه برای استخراج توضیحات دقیق‌تر عنوان مقالات کلمات را با استفاده از WordNet ریشه‌یابی کرده است. ماتریس کلمه–سند را با استفاده از خروجی بخش قبل تشکیل داده است. هر درایه در این ماتریس تکرار کلمه در سند مربوطه است.

تعداد چند عنوان را به‌صورت پیش‌فرض در محاسبه تعداد عنوان‌ها محاسبه کرده است و با استفاده از ترکیب LDA و Gibbs این عناوین را استخراج کرده است و برای هر عنوان چند کلمه *به‌دست‌آمده* است. با استفاده از کلمات *به‌دست‌آمده* برای هر عنوان بردار عنوان (بر اساس کلمه) *به‌دست‌آمده* است. برای مقایسه کردن بین سال‌های مختلف و سرطان‌های مختلف این بردارها نرمال شده‌اند (با استفاده از روش مین–مکس). میزان مشابهت بین بردارها نیز با استفاده از تابع Cosine محاسبه‌شده است (با این ضریب در تحلیل‌های پژوهش شباهت، میزان شباهت در موضوعات تحقیقات سرطان‌ها را با یکدیگر مورد ارزیابی قرار داده است). برای نتایج نیز بعد از موضوع‌بند و استخراج بردارهای موضوع‌ها آن‌ها را به‌صورت بصری بر روی نمودار (برای روند بر روی سال‌های

مختلف) نمایش داده است. میزان ترند تحقیقات را هم به‌صورت عددی در سال‌های مختلف مورد ارزیابی قرار داده است که چقدر تغییر کرده است.

۲-۲-۷- پیشینه پژوهش پیش‌بینی فناوری

همان‌طور که در ابتدای بخش بیان شد در پیش‌بینی فناوری، «فناوری» ابزارها، شیوه‌ها و فرایندهایی است که باهدف برآورده کردن نیازهای انسان مورداستفاده قرار می‌گیرد (Martino ۱۹۹۳). پیش‌بینی فناوری را به‌طورکلی هرگونه روش هدفمند و نظام‌مند برای پیش‌بینی و فهم در زمینه‌های نرخ رشد فناوری، جهت رشد فناوری، ویژگی‌های رشد آن و تأثیرات تغییرات فناوری (بخصوص ابداعات، نوآوری‌ها) خوانده‌اند (Firat, Woon, and Madnick۲۰۰۸). پیش‌بینی فناوری عبارت است از پیش‌بینی تمایلات و روند فناوری مانند جهت رشد و نرخ توسعه فناوری (Jun and Lee۲۰۱۲, Jun ۲۰۱۱, Porter ۱۹۹۱a).

درک و دانستن اینکه بخش تحقیق و توسعه چه منابعی در اختیار دارد و نیز مطالعه تحقیقات انجام‌شده، دو معیار مهم در رشد یک کسب‌وکار هستند. «مارتینو» در (Martino۲۰۰۳) مراحل تولید و توسعه فناوری را به بخش‌های زیر تقسیم کرده است:

- پیشنهادیه‌های نظری (تئوری)
- یافته‌های علمی
- امکان‌سنجی آزمایشگاهی
- نمونه اولیه
- معرفی تجاری

«فنگ-شانگ وو» نیز در ۲۰۱۱ نیز نوآوری‌های فنی را نیز به پنج گام دسته‌بندی کرده است (Wu et al.۲۰۱۱):

- تحقیقات اولیه
- تحقیقات کاربردی
- توسعه
- کاربرد

- تأثیرات اجتماعی

«روتولو» نیز در (Rotolo, Hicks, and Martin ۲۰۱۵) بیان کرده که رشد می‌تواند در جنبه‌های مختلفی انجام شود ازجمله:

- تعداد شرکت‌کنندگان در آن حوزه
- سرمایه‌گذاران خصوصی و عمومی
- تولیدات خروجی علمی
- نمونه‌های اولیه ارائه‌شده
- سرویس‌ها
- محصولات

ازآنجاکه با استفاده از دانش کشف‌شده از تحلیل‌روند می‌توان در برنامه‌ریزی استراتژی‌های آتی برای نوآوری‌ها فناوری استفاده کرد بنابراین می‌توان گفت که تحلیل‌روند تأثیر بسیار مهمی در پیش‌بینی فناوری دارد(.Wu et al ۲۰۱۱). یکی از مهم‌ترین منابعی که در کنار اختراعات ثبت‌شده در تحلیل‌روند (پیامد آن پیش‌بینی فناوری) مورداستفاده قرار می‌گیرد مقالات هستند (Wu et al. ۲۰۱۱).

در سال ۲۰۱۱ IARPA[۱] برای بودجه‌ای را برای پیش‌بینی و فهم تولیدات علمی در نظر گرفت که یکی از محورهای آن معرفی فناوری‌های نوظهور بود است (Wang ۲۰۱۸).

در (Ziegler ۲۰۰۹) و (.Newman et al ۲۰۱۴) از تحلیل‌روند و داده‌های ***مقالات*** برای پیش‌بینی فناوری استفاده کرده است.

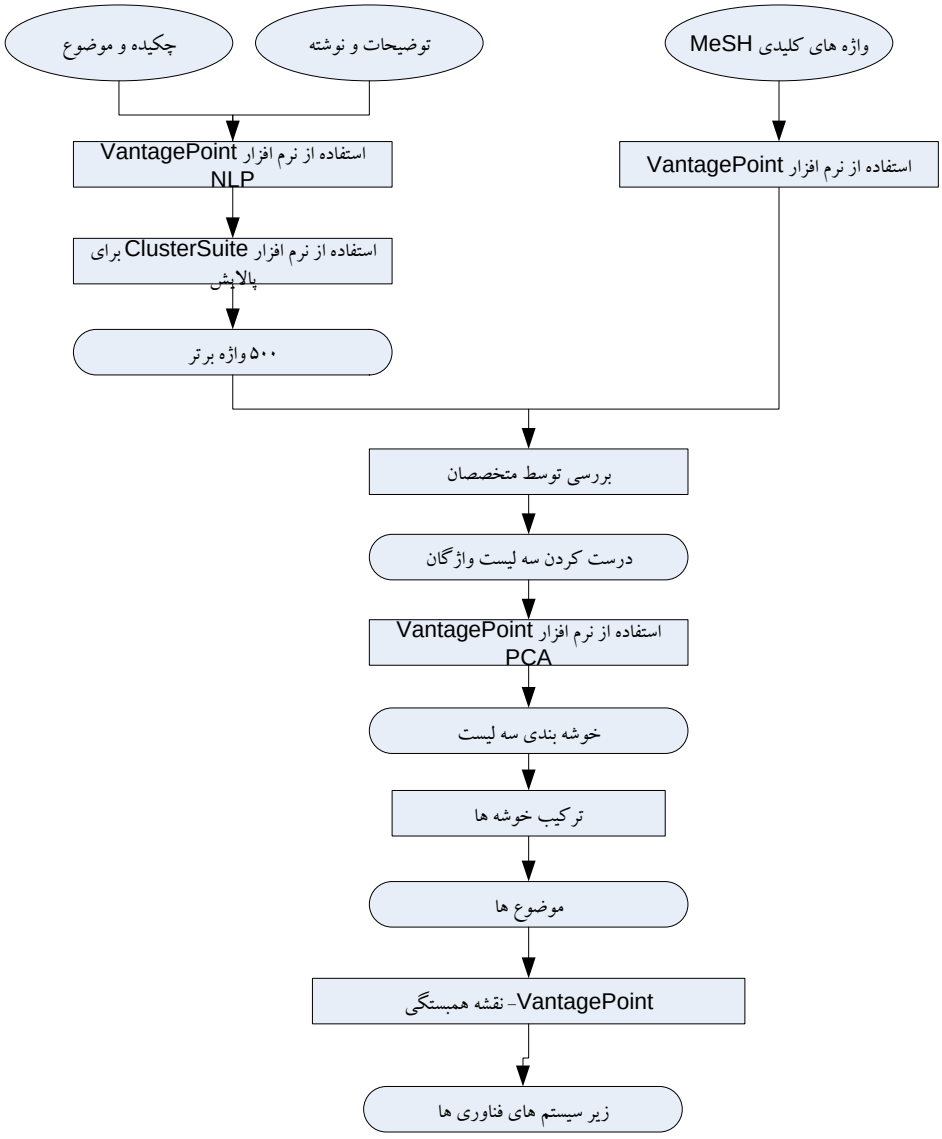
ازآنجاکه همه فناوری‌ها ممکن است از پیشرفت‌های علمی نشاءت نگیرند، لازم به ذکر است که در این رساله تمرکز تنها به پیش‌بینی فناوری‌هایی می‌پردازیم که قابلیت استخراج از اسناد علمی را دارند و قابلیت استخراج از اطلاعات کتابشناختی هستند.

[۱] Intelligence Advanced Research Projects Activity

رهیات کتابشناختی شامل محدودیت‌هایی چون تغییرات تصادفی پارساها مثلاً اضافه شدن دانشگاه‌های دیگر و افزایش چاپ شدن‌ها، تنها محدود بودن به منبع پارساها.

۲-۲-۷-۱- تحلیل موضوعی اسناد برای تشخیص مسیر فناوری و فرصت‌های بالقوه (Ma and Porter)

یک فرایند تحلیل موضوعی برای پیش‌بینی مسیر فناوری ارائه کرده است. بخش‌های این فرایند را در شکل ۲-۱۷ نشان داده‌شده است.



شکل ۲-۱۷ فرایند تحلیل موضوعی (Ma and Porter)

این فرایند برای ورودی خود سه منبع مختلف را دریافت می‌کند. منبع‌های ورودی عبارت‌اند از عنوان و چکیده، توضیح‌ها اضافی که به مستندات آن اختراع اضافه‌شده است و مجموعه کلمات‌کلیدی که از پایگاه داده‌ای مرتبط به سیستم اضافه می‌شود. از این منابع توسط نرم‌افزارهایی که از قبل طراحی‌شده بودند، کلمات‌کلیدی در سه لیست مختلف استخراج می‌شود.

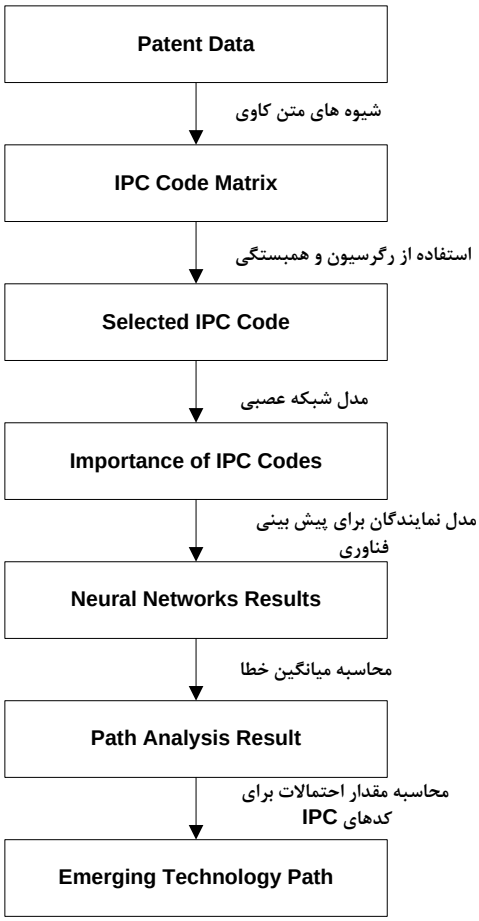
در فاز بعدی این فرایند سه لیست را در اختیار افراد متخصص قرار داده می‌شود که توسط این افراد اصلاح شوند. این قسمت از فاز کاری به‌صورت دستی انجام می‌شود که می‌تواند یک نقطه‌ضعف از سیستم محسوب شود. در دو مرحله بعدی بر روی هرکدام از فهرست‌های کلمات‌کلیدی، با استفاده از نرم‌افزار VantagePoint مبتنی بر روش تحلیل جزء اصلی، دسته‌بندی انجام می‌شود. خروجی این فاز سه مجموعه دسته مجزا بوده که هرکدام از این مجموعه‌ها بیانگر جنبه‌ها و ویژگی‌های جداگانه‌ای از موضوعه‌ای مطرح‌شده است. برای نمونه یک مجموعه مربوط به مزایا و معایب، مجموعه‌ای دیگر در مورد شیوه‌های دسترسی به آن موضوع و مجموعه بعدی مواد مورداستفاده برای دسترسی به یک داروی خاص است.

گام بعدی این سه مجموعه رده‌ترها را با یکدیگر ادغام کرده و به مجموعه جدید می‌رسد که بر اساس محتوای هرکدام از این دسته‌ها یک نام و عنوانی به آن دسته تخصیص داده می‌شود. با استفاده از نرم‌افزار ذکرشده و تعداد کلمات مشترک، ارتباط بین دسته‌ها را در یک‌شکل نشان می‌دهد. یک نمونه از این ارتباط را در شکل ۲ می‌توانید مشاهده کنید. حال که مجموعه‌ای دسته‌ها راداریم که هرکدام به فناوری و زمینه خاصی از موضوع اشاره می‌کند، می‌توانیم با رسم نمودار روند تکامل آن‌ها را در سال‌های مختلف موردبررسی قرار دهیم. مشاهده می‌شود که در بازه‌های زمانی چه فناوری‌هایی مورداستفاده قرارگرفته‌اند و به کدام سمت‌وسو حرکت می‌کنند.

۲-۲-۷-۲- پیش‌بینی فناوری‌های نوظهور با استفاده از تحلیل اطلاعات اختراع‌ها (Jun and Lee ۲۰۱۲):

در این فرایند داده ورودی سیستم، اطلاعات مربوط به اختراع‌ها است. از طریق روش‌های متن‌کاوی کدهای مربوط به هر اختراع را استخراج کرده که این اطلاعات کدهای IPC است که

نشان‌دهنده این است که اختراع زیرمجموعه کدام فناوری‌ها و دسته‌های ثبت‌شده است. سپس برای تمام اختراع‌های و لیست کدهای اختراع‌های مورداستفاده در آن‌ها، ماتریسی تشکیل می‌شود.



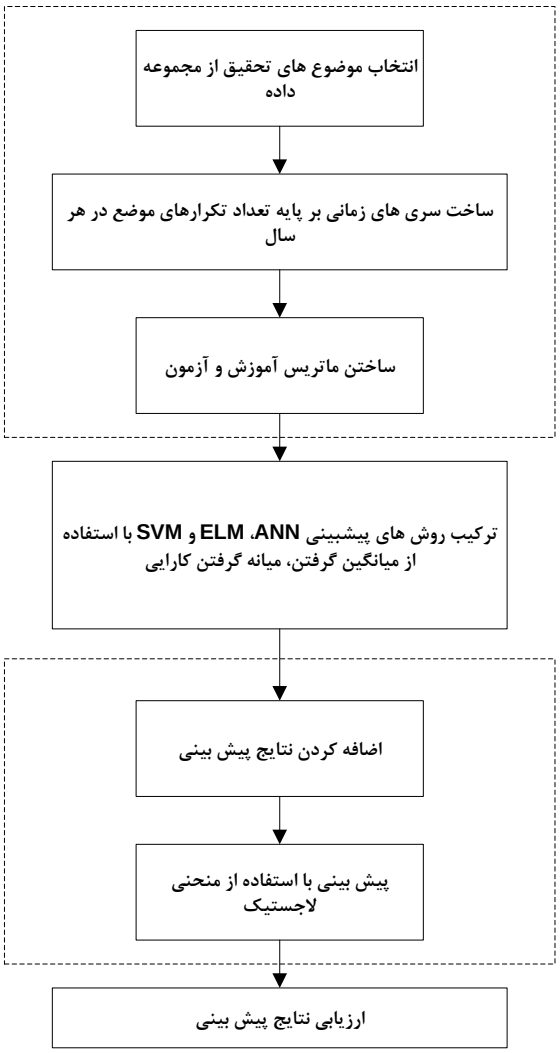
شکل ۲-۱۸فرایند مطرح‌شده

بعد از تشکیل ماتریس، کدهایی انتخاب می‌شوند که بیشترین رتبه را در بین بقیه دارند و سپس با استفاده از روش‌های رگرسیون و همبستگی کدهایی که اهمیت بیشتری را نسبت به بقیه دارند، را استخراج می‌کند. درنتیجه به یک مجموعه کدهای انتخاب‌شده خواهیم رسید که به فناوری هدف تأثیر بیشتری می‌گذارند. بعدازآن این مجموعه کدهای انتخاب‌شده را به یک شبکه عصبی داده می‌شود که بتواند فناوری نوظهور را با استفاده از داده‌های گذشته مدل کند. در این مدل ارائه‌شده، خروجی شبکه عصبی فناوری است که قرار است مورد پیش‌بینی قرار گیرد. تابع فعال‌ساز نرون‌ها تابع سیگموید و روش کاهش گرادیان نیز برای بهینه کردن وزن‌ها مورداستفاده قرار می‌گیرد. آن‌قدر

این عمل آموزش شبکه ادامه می‌یابد تا میانگین مربع خطا به حداقل برسد. از بهترین مدل *به‌دست‌آمده* می‌توانیم مسیر فناوری‌های نوظهور را بسازیم و از طریق نمودار آن را نشان می‌دهد.

۲-۲-۷-۳- پیش‌بینی موضوع‌های علمی با استفاده از ترکیب یادگیری ماشین و منحنی لجیستیک (AGUS WIDODO۲۰۱۳)

در این روش مجموعه‌داده‌های ورودی از گزارش‌های تحقیقاتی تهیه‌شده هستند. این مجموعه‌داده توسط ۵۰ کلمه کلیدی معروف گروه‌بندی می‌شوند. در گام بعدی این گروه‌ها را به ترتیب زمان مرتب‌شده و نمودارهای سری زمانی را برای هرکدام تشکیل داده می‌شود. بر اساس آن ماتریس‌هایی برای آموزش ماشین مرحله بعدی تشکیل‌شده است. نکته قابل‌ذکر این است که کلمات‌کلیدی از قبل مشخص هستند و در ضمن هر گزارش در یک گروه بیشتر طبقه‌بندی نشده است که هر دو از معایب این فرایند هستند. شکل ۴ بلاک‌های این فرایند را نشان می‌دهد.



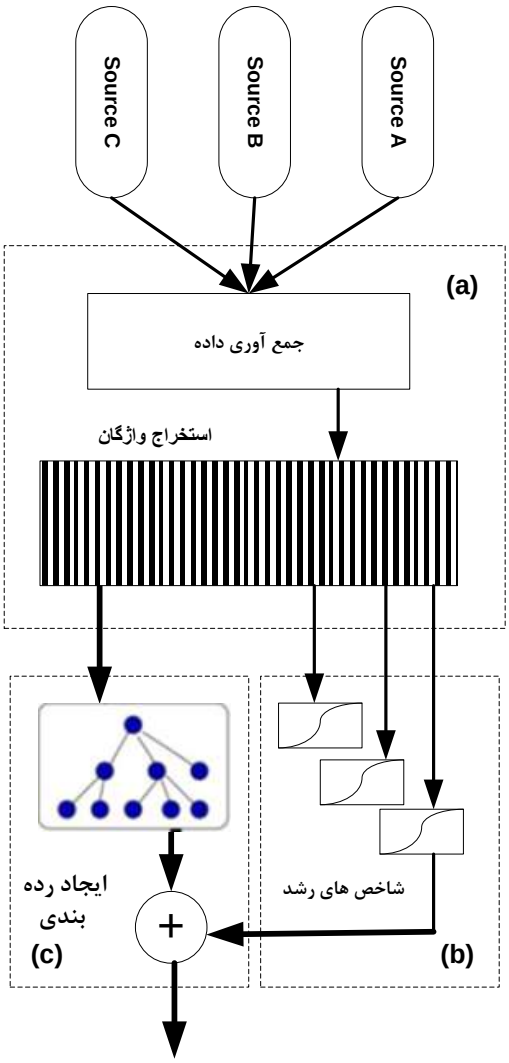
شکل ۲-۱۹ فرایند پیش‌بینی آینده ارائه‌شده در [۸۷]

در مرحله بعدی برای پیش‌بینی از سه روش مختلف یادگیری ماشین; شبکه عصبی مصنوعی، روش ماشین بردار پشتیبان (SVM) و یادگیری ماشینی شدید (ELM). یکی از خروجی‌های این بخش مقایسه کارایی این سه ماشین در سرعت و دقت پیش‌بینی در هست. بعد از آموزش این سه ماشین، برای کارایی بهتر خروجی این سه ماشین را میانگین گرفته است و سپس برای تنظیم کردن بیشتر به منحنی S شکل می‌دهند. همان‌طور که ملاحظه کردید در این الگوریتم نویسنده از منحنی لجستیک برای پیش‌بینی استفاده کرده است. از این منحنی در کارهای دیگر برای آینده‌نگری فناوری مورداستفاده شده است (Daim et al. ۲۰۰۶, Smalheiser ۲۰۰۱, Christodoulos, Michalakelis, and Varoutas ۲۰۱۰).

۲-۲-۷-۴- آینده‌نگری فناوری اطلاعات با استفاده از داده‌کاوی و معناشناختی

(Firat, Woon, and Madnick ۲۰۰۸) (Vidican, Woon, and Madnick ۲۰۰۹, Woon, Zeineldin, and Madnick ۲۰۱۱):

دانشگاه MIT پروژه‌ای را تحت عنوان پیش‌بینی آینده در سال ۲۰۰۷ تا ۲۰۱۱ (با همکاری ۱۰نفر از استاید و ۳نفر از دانشجویان) انجام داده که در آن قالب کاری را برای پیش‌بینی آینده فناوری ارائه کرده است و علاوه بر قالب کاری از این طرح آن چندین مقاله در رابطه با تحلیل‌روند و پیش‌بینی فناوری (Dawelbait et al. ۲۰۱۱, Henschel et al. ۲۰۱۲, Dawelbait et al. ۲۰۱۰) تحلیل کتابشناختی (Vidican, Woon, Henschel, and Madnick ۲۰۰۹)، و شیوه‌های طبقه‌بندی برای ارزیابی (Woon and Madnick ۲۰۰۹, Dawelbait et al. ۲۰۱۰, Henschel et al. ۲۰۰۹) ارائه‌شده است. این قالب کاری در شکل ۴ نشان داده‌شده است که در آن می‌توان از منابع مختلف به‌عنوان ورودی‌ها برای سیستم استفاده شود که این دانشگاه منبع خود را مقالاتی که در SCOPUS آمده است را به سیستم خود اعمال کرده و از طریق روش‌های کتابشناختی برای تحلیل بهره برده است.



شکل ۲-۲۰ قالب کاری ارائه‌شده

همان‌طور که ذکر شد از طریق روش‌های کتاب‌سنجی از منابع ورودی سیستم کلمات‌کلیدی که مرتبط با حوزه موردبررسی هستند را استخراج می‌شود. لازم به ذکر است این قسمت بسیار مهم است زیرا تمام اطلاعات و تحلیل‌های بعدی (در رابطه با فناوری آینده و جهت تحقیقاتی) وابسته به کیفیت کلمات استخراج‌شده دارد. در گام بعدی بر اساس فاصله معنایی و روش‌های خودسازمان‌دهی، به نحوی آینده فناوری را ترسیم کند.

قسمت دوم این قالب کاری، سه شاخص رشد فناوری معرفی کرده تا بر اساس آن به محاسبه میزان رشد هریک از فناوری‌هایی است که کلمات‌کلیدی مرتبط با آن نشان می‌دهد بپردازد. اینکه شاخصه‌ای رشد چگونه انتخاب شوند و کدام شاخص رشد اهمیت بیشتری را دارد از اهمیت ویژه‌ای

برخورداراست. در قسمت سوم کلمات‌کلیدی را به نحوی باهم مرتبط کرده تا مجموعه کلمات‌کلیدی یک فناوری خاصی را ارائه دهند. این طبقه‌بندی برای این است که هریک از کلمات‌کلیدی به‌تنهایی می‌توانند دارای نویز باشند و به این خطا در پیش‌بینی ما اثر کمتری داشته باشند.

۲-۲-۸- روش‌های ارزیابی در کشف موضوع‌ها

روش ارزیابی‌های ارزیابی استخراج‌شده از مطالعات انجام‌شده به شرح زیر است:

۱. استفاده از **متخصصان** (که آیا موضوعات موردعلاقه هستند، یا موضوعات را تائید می‌کنند) (این امکان به خاطر بایاس فکری، گسترده بودن یک حوزه علمی و عدم شناخت حوزه‌های جدید این مسئله نیز با مشکل مواجه است. البته بازهم به خاطر تعاریف مختلف از موضوع و موضوع‌های جدید این ارزیابی بازهم با مشکل مواجه می‌شود (Wang ۲۰۱۸).) : (Tu and Seng ۲۰۱۲) و (Carley et al. ۲۰۱۸) , (Zhang et al. ۲۰۱۸)

۲. استفاده از یک مثال برای نمایش عملکرد سیستم: (Tu and Seng ۲۰۱۲) و (Wang ۲۰۱۸)

۳. مقایسه اجزای طرح که بهبود داده‌اند (مثلاً خوشه‌بندی) : (Zhang et al. ۲۰۱۸)

۴. مشاهده شواهد موجود - بررسی نظری نتایج موجود در اینترنت یا روش‌های قبلی که با روش‌های متفاوت و داده‌های متفاوت و در بازه‌های زمانی متفاوت انجام‌شده و مقایسه نتایج آن‌ها (که فقط بررسی کنند که عملیات حدودا درست‌کار می‌کند یا نه در (Tu and Seng ۲۰۱۲) (Wang ۲۰۱۸) و (Small, Boyack, and Klavans ۲۰۱۴)

۵. روش باید ویژگی‌های مستقیم، قابل‌درک، شفاف بودن را داشته باشد (Wang ۲۰۱۸)

۶. از داده‌های با برچسب استفاده کند (Kasiviswanathan et al. ۲۰۱۱).

به خاطر عدم وجود استاندارد و معیار فیزیکی دقیق، روشی برای ارزیابی موضوع (یا کلمه) استخراج‌شده یا طبقه‌بندی آن‌ها وجود ندارد (Klavans and Boyack ۲۰۱۷). عدم وجود معیار شناخته‌شده‌ای برای ارزیابی موضوع‌های جدید می‌تواند به خاطر عدم وجود فهرستی از موضوع‌های جدید باشد (Small, Boyack, and Klavans ۲۰۱۴). مقایسه باکارهای گذشته دلیلی بر بهبود در

ارزیابی یک طرح نیست (Wang۲۰۱۸). دو دلیل برای عدم امکان بررسی باکارهای گذشته وجود دارد: خصوصیات و تعریف موضوع‌ها (بخصوص موضوع‌های جدید) با یکدیگر متفاوت هستند، دوم توافق نظری بر روی تعریف و خصوصیات نیست. در این پژوهش علاوه براینکه روش ارائه‌شده دادای ویژگی‌های مستقیم، قابل‌درک، شفاف بودن است، اجزای سیستم ارائه‌شده را با دیگر روش‌ها مقایسه خواهیم کرد و با نمایش یک مثال عملکرد آن را بررسی خواهیم کرد و در برخی از اجزای سیستم ارائه‌شده از داده‌های برچسب‌گذاری شده استفاده خواهیم کرد.

جدول ۲-۱خلاصه برخی روش‌های تحلیل‌روند و فناوری

مقاله	داده‌ها	انتخاب کلیدواژه	بازنمایی کلمات	موضوع‌بندی	تحلیل روند
An and Wu (۲۰۱۱)	مقالات	آنتروپی و نظر متخصصان	همرخدادی – هر مقاله یک هم‌رخدادی	سلسله‌مراتبی	نمودار راهبردی
MIT ۲۰۰۹-۲۰۰۱۱	مقالات	تعداد کلیدواژه‌های ثابت	ندارد	بر اساس روابط معنای ویکیپدیا رده‌بندی می‌کند	سری‌زمانی – شاخص رشد نسبت به سال گذشته
Wu et al. (۲۰۱۱)	مقالات- اختراعات	۵۰ کلیدواژه برتر	همرخدادی – هر مقاله یک هم‌رخدادی	دارد	جدول بازه‌های زمانی – سری زمانی موضوع‌ها
Hu et al. (۲۰۱۳)	مقالات	کلیدواژه‌ها با تکرار مشخص	همرخدادی – هر مقاله یک هم‌رخدادی	سلسله‌مراتبی	نمودار راهبردی
Jun and Lee (۲۰۱۲)	ثبت اختراع	کدهای اختراع (نمایه‌ها مشخص کننده حوزه اختراع)	ندارد (تنها فراوانی کلمه)	ندارد	نمودار سری زمانی – پیش‌بینی با شبکه عصبی

AGUS) WIDODO (۲۰۱۳	مقالات	تعداد کلیدواژه‌های ثابت	ندارد (تنها فراوانی کلمه)	ندارد	نمودار سری زمانی کلمات- پیش‌بینی با شبکه عصبی
Buscaino) ۲۰۱۴	مقالات	۱۰ کلمه‌ای که بیشترین تغییر را داشتند.	W2V	دارد- pca-mds	سری زمانی موضوع‌ها
Ojo and) Adeyemo ۲۰۱۵	مقالات	انتخاب بر اساس TFو DF	VSM-TF	موضوع‌بندی دستی	سری زمانی موضوع‌ها
Hue 2015	مقالات-	کلیدواژه‌ها با تکرار مشخص	همرخدادی - هر مقاله یک هم‌رخدادی	سلسله‌مراتبی	نمودار راهبردی- شبکه‌های اجتماعی
Ma and) ۲۰۱۵ (Porter	مقالات	انتخاب ۵۰۰ کلیدواژه برتر (دستی)	VSM	دارد (مشخص نکرده)	سری زمانی موضوع‌ها
Kim et al.) ۲۰۱۷).	مقالات – پروژه‌های پژوهشی	انتخاب محدود کلمات	تشکیل آنتولوژی با استفاده از خبرگان	گلیدواژه‌های نزدیک به صورت دستی خوشه (یک موضوع) می‌شوند	سری زمانی موضوع‌ها – اطلاعات آماری –شبکه کلمات
Tao et al.) (۲۰۱۷	مقالات	انتخاب دستی کلیدواژه‌ها	W2V	کا-میانه	سری زمانی موضوع‌ها
سهیلی ۱۳۹۷	مقالات	انتخاب دستی کلیدواژه‌ها	همرخدادی - هر مقاله یک هم‌رخدادی	سلسله‌مراتبی	نمودار استراتژیک

Chang et al. (۲۰۱۷)	مقالات	کلیدواژه‌ها بر اساس تعداد تکرارهای یکی	همرخدادی – هر مقاله یک هم‌رخدادی – ماتریس همبستگی	سلسله مراتبی	آماری
Jiang et al. ۲۰۱۸	مقالات	استخراج دستی عبارات مهم – امتیازدهی دستی بر اساس نویسندگان			سری زمانی
(Wu et al. ۲۰۱۸)	مقالات	تعداد کلیدواژه‌های ثابت	-	LDA	سری زمانی موضوع‌ها
Zhao et al. (۲۰۱۸)	مقالات	تعداد کلیدواژه‌های با تکرار بالاتر از ۲۰	هر مقاله یک هم‌رخدادی	سلسله مراتبی	نمودار راهبردی – شبکه‌های اجتماعی

فصل ۳- رویکرد پیشنهادی

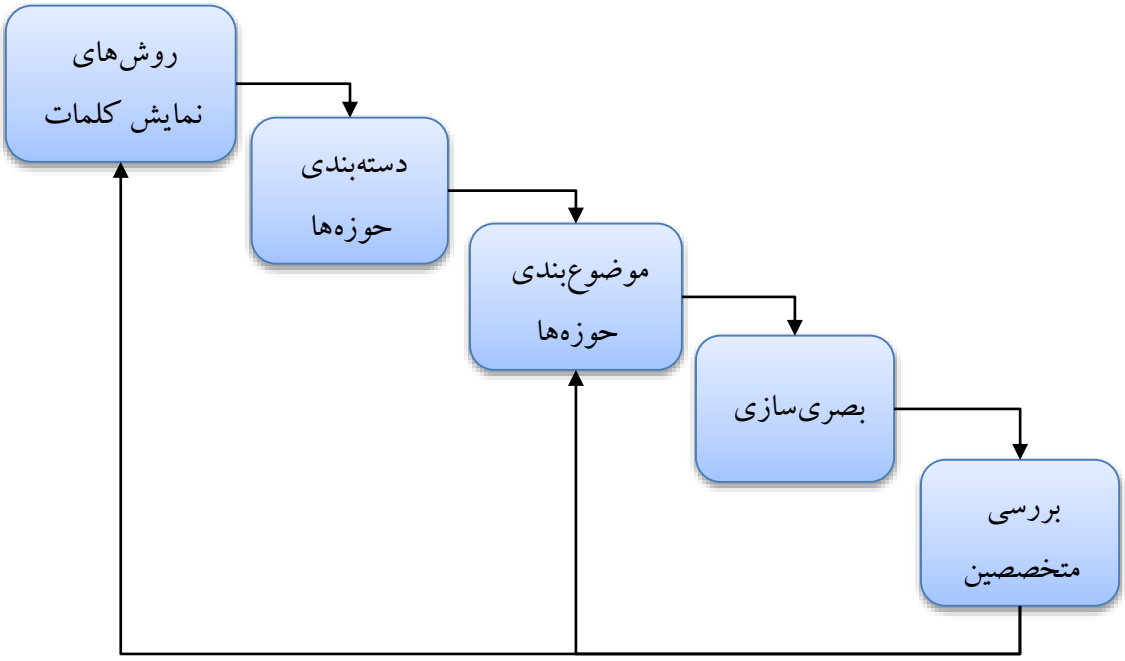
در این فصل، روش ارائه‌شده برای تحلیل‌روند در این رساله را بیان می‌شود. ساختار این فصل را می‌توان به‌صورت ذیل خلاصه نمود: در ابتدا به بیان پیش‌زمینه‌های موردنیاز پرداخته و سعی در فهم بهتر مسئله خواهیم داشت. سپس با نگاهی دقیق‌تر به بررسی داده‌ها می‌پردازیم. گام بعدی بیان چگونگی آماده‌سازی و پیش‌پردازش داده‌ها خواهد بود. سپس روش‌هایی برای موضوع‌بندی حوزه‌های علمی ارائه و موردبحث و بررسی قرار خواهد گرفت. با توجه به خروجی‌های این روش در گام بعد (بخش۴) با استفاده از نمودار راهبردی (و ارائه یک روش محاسبه میزان چگالی و مرکزیت) روند آن حوزه علمی و موضوع‌های آن حوزه را موردبررسی، تحلیل و نمایش قرار خواهیم داد. درنهایت موضوع‌های استخراج‌شده تحلیل‌روند خواهند شد. در هرکدام از مراحل ذکرشده، روش‌های پیشنهادی برای بهینه کردن آن بخش ارائه خواهد شد.

۳-۱- روش پژوهش (ساختار پژوهش)

با استفاده از کشف و تشخیص موضوع و سپس دنبال کردن آن در طول یک بازه زمانی می‌توان روند آن موضوع را موردبررسی قرار داد. برای مثال، رسم تعداد اسناد چاپ‌شده در یک موضوع مشخص در طول زمان، روند تکامل یک موضوع در طول آن زمان را به ما نشان می‌دهد (Rajaraman and Tan ۲۰۰۱). هدف اصلی در این رساله تحلیل‌روند یک حوزه‌علمی است (برای توضیحات بیشتر در خصوص تحلیل‌روند و روش‌های پیش‌بینی فناوری به فصل دوم مراجعه شود) به همین منظور نیاز است ابتدا موضوع‌های موجود در یک حوزه مشخص شود و سپس آن موضوع‌ها تحلیل و ارزیابی گردد. از آنجا که تحلیل‌روند قرار است با استفاده از روش‌های داده‌کاوی انجام شود، روش انجام این پژوهش و پیاده‌سازی آن در این رساله مطابق استاندارد CRISP-DM صورت گرفته‌است (Shearer ۲۰۰۰). در شکل ۲-۴ روش پژوهش و استانداردCRISP-DM **Error!**

Reference source not found. که به طور کلی برای انجام این پژوهش مورد استفاده قرار گرفته است.

در ادامه این فصل به بررسی دقیق‌تر روش‌پیشنهادی برای تحلیل‌روند بر اساس گام‌های استاندارد CRISP-DM خواهیم پرداخت. در فصل اول به بیان و فهم مسئله و اهداف رساله پرداخته شده است لذا برای تکرار نشدن مطالب به فصل اول مراجعه نمایید. از آنجا که تحلیل‌روند می‌تواند براساس اهداف مختلف، با داده‌های مختلفی انجام شود و داده‌های در اختیار ما پارسای موجود در پایگاه گنج پژوهشگاه علوم و فناوری اطلاعات ایران است، لذا در بخش دوم به بررسی اسناد موجود خواهیم پرداخت و با خصوصیات آن‌ها آشنا خواهیم شد. . با استفاده از نوع داده‌های موجود در بخش سوم این فصل با توجه به نیازمندی‌های پژوهش و داده‌های موجود، بعد از پیش‌پردازش‌های اولیه، مجموعه‌داده‌هایی را برای انجام تحلیل‌های آتی نیاز است را آماده خواهیم کرد. در بخش بعدی روش‌هایی برای موضوع‌بندی حوزه‌های علمی ارائه خواهیم کرد و با توجه به خروجی‌های این روش در گام بعد (بخش ۳-۵-) با استفاده از نمودار راهبردی و ارائه یک روش برای تحلیل روابط موضوعات کشف شده، روند آن حوزه علمی و موضوع‌های آن حوزه را موردبررسی، تحلیل و نمایش قرار خواهیم داد. برای افزایش دقت و کارایی و کاهش خطاهای ابزارهای مورد استفاده، متخصصان حوزه موردتحلیل می‌توانند برخی حوزه‌ها را حذف یا ادغام کنند. پیچیده‌بودن گام مدل‌سازی این گام را به اجزای کوچکتر تقسیم کرده و به‌صورت زیر ارائه‌شده است.



شکل ۳-۱ گام‌های مختلف برای مدل‌سازی داده‌ها در رویکرد پیشنهادی

۳-۲- شناخت داده‌ها

هدف اصلی این رساله تحلیل‌روند و پیش‌بینی فناوری یک حوزه علمی است که در زمینه‌های مختلفی ازجمله سیاست‌گذاری، تعیین نقشه چشم‌انداز، آگاه‌سازی و تولید نقشه راه محققین نقش مهمی دارد. بافهم و درک روند یک حوزه علمی و پیدا کردن زیر حوزه‌های آن می‌توان میزان پیشرفت در آن حوزه را با وضعیت رشد منطقه و یا جهانی مقایسه کرد. برای تحلیل‌روند از داده‌های مختلفی از جمله مقالات و اسناد علمی، اختراعات چاپ شده اسناد می‌توان استفاده کرد (برای مطالعه بیشتر به بخش ۲-۲-۶- و بخش ۲-۲-۷- مراجعه شود).

داده‌هایی که در این پژوهش مورداستفاده قرار می‌گیرند، پارساهای موجود در پایگاه داده گنج پژوهشگاه علوم و فناوری اطلاعات ایران است. از آنجا که اسناد مختلف ساختار و خصوصیات نوشتاری مختلفی دارند لذا می‌بایست پارسای در دسترس را مورد بررسی بیشتری قرار دهیم تا با خصوصیات نوشتاری آن‌ها بیشتر آشنا شویم. برای مثال با شناخت ساختار داخلی چکیده‌های پارسا می‌توان میزان جامعیت نمایه‌های انتخاب شده را مورد بررسی قرار داد.

بنابراین هدف اصلی این گام از پژوهش بررسی و تحلیل این اطلاعات آماری است تا علاوه بر اینکه معایب در انتخاب نمایه و نگارش چکیده کشف شود، اطلاعاتی نیز در خصوص نحوه انتخاب کلیدواژه‌ها ارائه شود. از آنجاییکه یکی از مهمترین اقدامات برای تحلیل خودکار مستندات علم، استخراج کلیدواژه‌های مناسب است، نحوه انتخاب کلیدواژه‌ها یکی از گام‌های اصلی برای تحلیل روند محسوب می‌شود. مهم‌ترین سؤالات این بخش از پژوهش عبارت‌اند از:

- رابطه و الگوی کلیدواژه‌های انتخابی نویسنده و نمایه‌ساز چگونه است؟
- چه میزان کلیدواژه‌های انتخابی نویسنده در عنوان و چه میزان در چکیده و عنوان پارسا قرار دارند؟
- توزیع کلیدواژه‌ها در چکیده به چه صورت است؟
- آیا چکیده نویسی پارسای موجود مطابق با استاندارد است؟

ازآنجاکه دسترسی به متن کامل این پارساها مقدور نبود، تحلیل‌ها بر روی فراداده‌های آن‌ها انجام خواهد شد:

- عنوان پارسا؛
- چکیده پارسا؛
- کلیدواژه‌های نویسنده؛
- کلیدواژه‌های نمایه‌ساز؛
- دانشگاه محل تحصیل؛
- دانشکده محل تحصیل.

مطابق با پژوهش (He ۱۹۹۹) و بیان (Lv et al. ۲۰۱۱) نویسندگان اسناد علمی سعی می‌کنند مطالب اصلی را در عنوان سند خود بیان کنند. بنابراین می‌توان برای کشف نمایه‌های یک سند از کلمات عنوان آن سند استفاده کرد. مهم‌ترین تحلیل‌های بکار گرفته‌شده در این بخش شامل موارد زیر هستند:

- میزان اشتراک نمایه‌های نویسنده و نمایه‌ساز؛
- میزان کلیدواژه‌های انتخابی نویسنده در عنوان و در چکیده؛
- توزیع نمایه‌های استفاده‌شده در عنوان و چکیده؛
- بررسی چکیده نویسی پارسای موجود و انطباق آن‌ها با استاندارد ایزو.

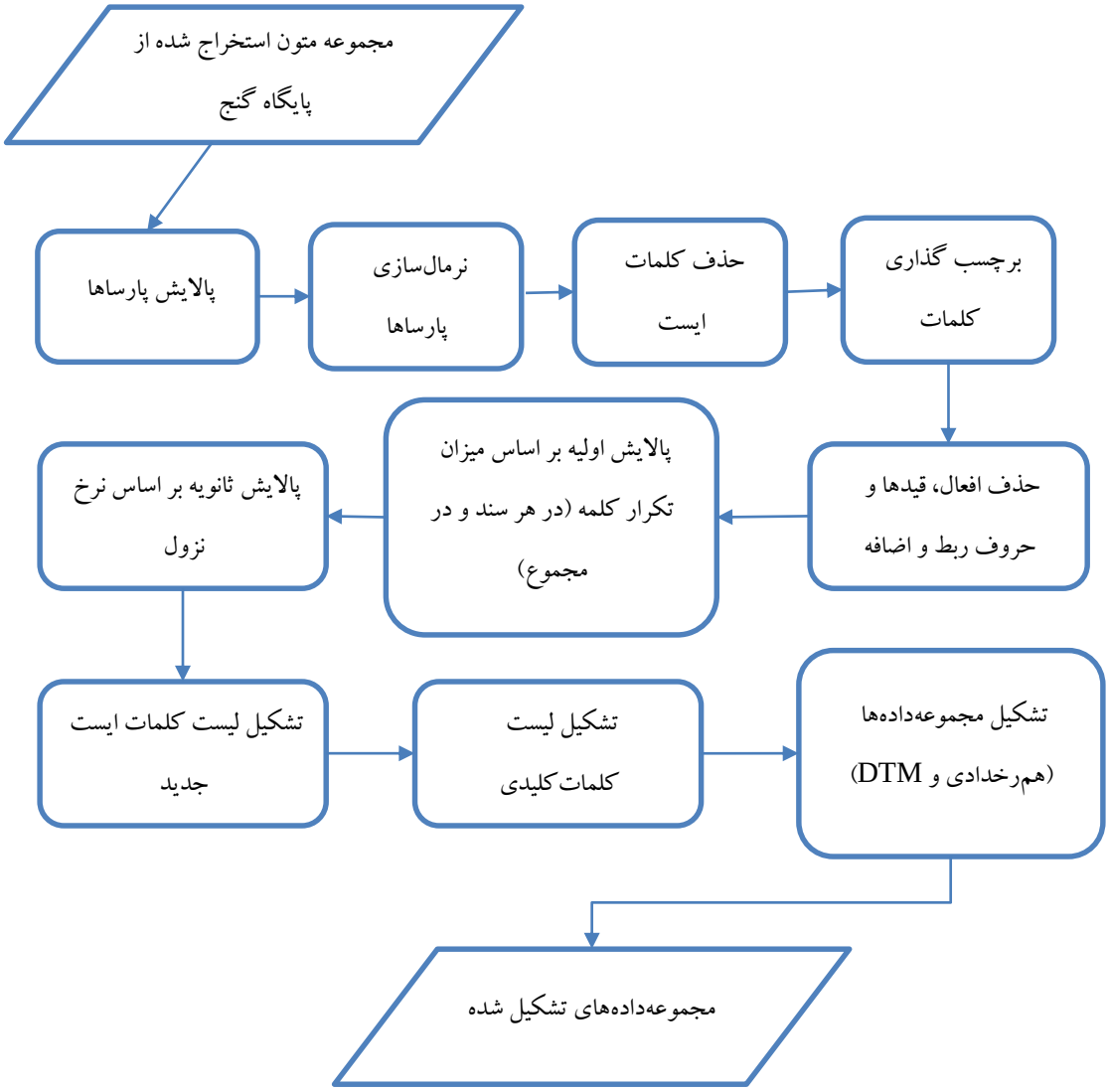
در انتهای این بخش بعد از استخراج میزان اشتراک انواع نمایه‌ها، با استفاده از نحوه میزان توزیع نمایه‌ها در عنوان و چکیده توانایی وزن‌دهی به عنوان و چکیده را خواهیم داشت. با استفاده از این امر قابلیت استخراج خودکار با درصد درستی بالاتر نمایه‌ها را خواهد شد. با مقایسه چکیده‌ها با استاندارد ایزو چکیده نویسی قابلیت افزایش بهبود کیفیت نوشتاری چکیده و انتخاب کلیدواژه خواهد بود.

۳-۳- آماده‌سازی داده‌ها

بعد از شناخت خصوصیت داده‌ها، می‌بایست آن‌ها را به شکلی تبدیل کنیم که قابل پرداش برای رایانه باشد. بنابراین در مرحله جاری آماده‌سازی داده‌ها برای اعمال پردازش رایانه‌ای انجام می گیرد[۱]. هر سندی از مجموعه کلمات تشکیل شده است که می‌بایست کلمات آن‌ها را به شکل قابل پردازش برای رایانه آماده کنیم. در بخش ۲-۱-۷- برخی روش‌های تبدیل کلمه به شکل قابل پردازش توسط رایانه آمده است. اشکال مختلف بازنمایی کلمات (نحوه نمایش کلمات) در کارایی یک سیستم تاثیر مستقیم می گذارد.

از آنجا که مجموعه کلیدواژه‌های یکتای متون بسیار زیاد است در این مرحله بعد از استخراج کلیدواژه‌های کاندید با استفاده از روش‌های معمول روشی برای پالایش و انتخاب کلیدواژه‌های با تاثیر بیشتر ارائه‌شده است. مراحل انجام این گام در شکل ۳-۲ آمده است. *سهم علمی دیگری که در این بخش ارائه‌شده است، روشی برای بازنمایی کلمات (نمایش کلمات) و تشکیل ماتریس هم‌رخدادی با هدف بهبود در موضوع‌بندی حوزه‌های علمی و به تبع آن تحلیل‌روند بهتر است.*

[۱] همانطور که قبلا نیز بیان شد داده‌های این پژوهش فرادادهای پارسا ی موجود در پایگاه داده گنج است.



شکل ۳-۲مراحل آماده‌سازی مجموعه‌داده‌ها

قبل از هر چیز از مجموعه داده دریافت شده، رکوردهایی که فیلدهای ذکر شده در بخش قبل را ندارند را از پردازش‌های بعدی خود حذف می‌کنیم. در مرحله‌های بعدی پیش‌پردازش بر روی داده‌های نوشتاری[۱] مورداستفاده انجام می‌گیرد. در ابتدا بر روی تمام متون پیش‌پردازش هنجارسازی نوشتار (متن)[۲] انجام گردیده است.

[۱] Text data

[۲] Text normalization

مطالعات نشان داده‌اند که با استفاده از کلیدواژه‌های عنوان و کلیدواژه‌های نویسنده می‌توان به هدف، نوآوری و موضوع اصلی متن پی برده شود. توضیحات تکمیلی‌تر در مقاله‌ای که در این خصوص به چاپ رسانده‌ایم در دسترس است. (Khatir and Ganjefar ۲۰۱۶, He ۱۹۹۹). بنابراین علاوه بر کلیدواژه‌های عنوان کلیدواژه‌های مهم استفاده‌شده در اطلاعات کتاب‌شناختی استخراج به کلیدواژه‌ها اضافه شده‌اند. این کلیدواژه‌ها با استفاده از مراحل بعدی پالایش خواهند شد.

بعد از هنجارسازی کلمات، نوبت به کلمات و حذف کلمات‌ایست می‌رسد.کلمات‌ایست به کلماتی گفته می‌شوند که مفهومی را منتقل نمی‌کنند و بار معنایی خاصی در آن موضوع ندارند. کلمات ایست[۱] بر دو دسته تقسیم می‌شوند:

- کلمات‌ایست در یک زبان
- کلمات‌ایست در یک حوزه.

برای حذف کلمات‌ایست عمومی یک زبان از لیست‌های آماده می‌توان استفاده کرد (که امری متداول و ساده است). اما برای حذف کلمات‌ایست یک حوزه نیاز به بررسی بیشتر آن حوزه توسط کارشناسان آن حوزه یا استفاده از نتایج پردازش زبانی است که در آن حوزه انجام‌شده است. در این پژوهش ما علاوه بر حذف کلمات‌ایست عمومی در هر پردازش، یک حوزه خاص لیست کلمات عمومی آن حوزه را نیز استخراج خواهیم کرد تا در آینده برای پردازش یک پارسا مرتبط با آن حوزه بتوانیم کلمات عمومی آن حوزه را برای استخراج کلمات مفیدتر از آن سند حذف کنیم. لیست کلمات‌ایست عمومی مورداستفاده در این رساله از (Davarpanah, Sanji, and Aramideh ۲۰۰۹) استفاده‌شده است.

برای تشکیل کلمات‌ایست یک حوزه، بعد از تشکیل لیست کلمات باقی‌مانده (بعد از حذف کلمات‌ایست عمومی) و سپس هنجارسازی کردن کلمات، ریشه کلمات را استخراج کرده و آن‌ها

[۱] Stop words

را برچسب اجزای کلامی[1](POS)[2] می‌زنیم و کلماتی که در موقعیت فعل قرار دارند را حذف می‌کنیم. کلمات بااهمیت در این رساله کلماتی هستند که از اجزای اسم، صفت و قید تشکیل‌شده باشند. بنابراین با پالایش کردن کلمات با استفاده از POS کلمات بااهمیت بیشتر را جدا خواهیم کرد. برای این منظور از توابع برچسب‌گذاری و استخراج ریشه کلمات که توسط آزمایشگاه دانشگاه فردوسی مشهد طراحی‌شده در این رساله مورداستفاده قرارگرفته است.

فهرستی از ریشه کلمات استخراج‌شده است (که شامل قید، صفت و اسم هستند) و در گام بعد میزان تکرار این کلمات را درمجموعه‌ای فرا داده‌ای که در اختیارداریم محاسبه کرده و غیر از محاسبه تکرار آن‌ها معیارهای زیر را در آن‌ها محاسبه خواهیم کرد:

- میزان تکرار کلمه[3]
- میزان تکرار سندهایی که در آن کلمه تکرار شده است[4]
- TF*IDF

در مرحله بعدی پالایشی بر اساس سه پارامتر فوق انجام می‌شود و کلمات با تکرار پایین و تکرار بسیار زیاد حذف خواهند شد.کلمات دارای تکرار پایین از نظر معنایی روش‌های بازنمایی کلمات را با خطا می‌توانند مواجه کنند (Li and Wang۲۰۱۷) بنابراین نیاز به حذف آن‌ها هستیم. کلمات با تکرار بالا در این رساله به کلماتی گفته می‌شود که که در بیشتر از نصف اسناد آمده‌اند. کلمات با تکرار پایین کلماتی هستندکه در کمتر از ۵ درصد از رساله‌ها آمده‌اند. در دو مجموعه کلمات به مجموعه کلمات‌ایست اضافه می‌شوند.

---

[1] برچسب «اجزای کلامی» یا «بخش گفتاری» به گروهی از کلمات زده می‌شود که در مکان و ساختار مشابهی در جمله از نظر دستور زبانی قرار می‌گیرند. برای مثال «اسم خاص»، «فعل» یا «فعل کمکی» از بر چسب‌های اجزای کلامی محسوب می‌شوند.

[2] Part Of Speech (POS)

[3] Term Frequency (TF)

[4] Document Frequency (DF)

در مرحله پالایش ثانویه، روشی ارائه‌شده است تا کلماتی که تاثیر بیشتری نسبت به کلمات باقیمانده دارند انتخاب شوند. برای این منظور روابط (۳-۱) تا (۳-۴) ارائه‌شده است. این رابطه بر این اصل طراحی‌شده است تا به صورت خودکار آن دسته از کلماتی که تمایز زیادی نسبت به ندارند حذف شوند. این کلمات دو ویژگی اصلی دارند:

- تعداد زیادی از کلمات دیگر هستند که فراوانی تقریباً یکسانی نسبت به این کلمات دارند.
- نرخ تغییر فراوانی این کلمات با کلمات بااهمیت‌تر زیاد، و فراوانی آن‌ها کمتر است.

$$Sort\ TFs \tag{۱-۳}$$

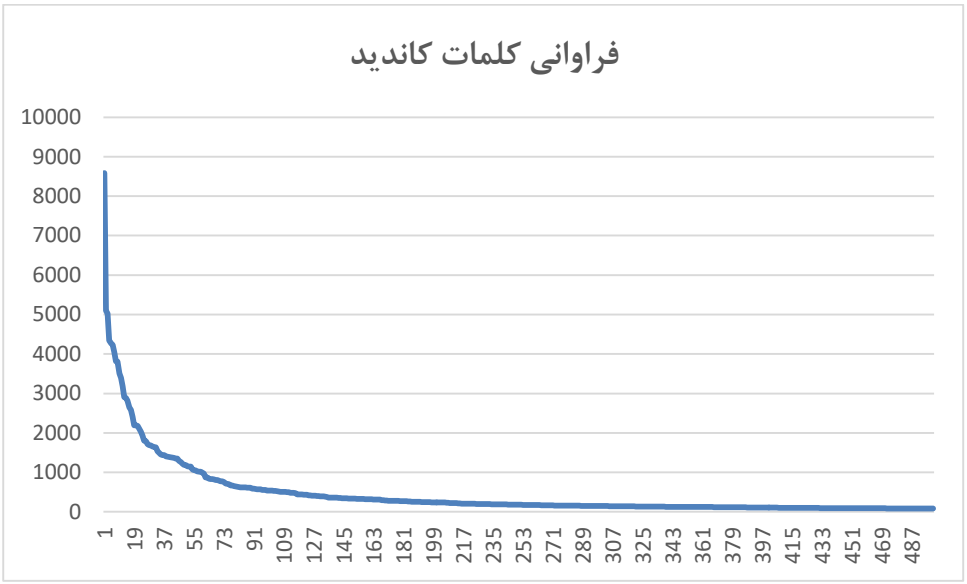
$$v_i = \frac{۲ * tf_i - tf_{i+۱} - tf_{i+۲}}{۲} \tag{۲-۳}$$

$$|av| = \sum_{j=i-m}^{i+m} \frac{v_i}{۲m + ۱} \tag{۳-۳}$$

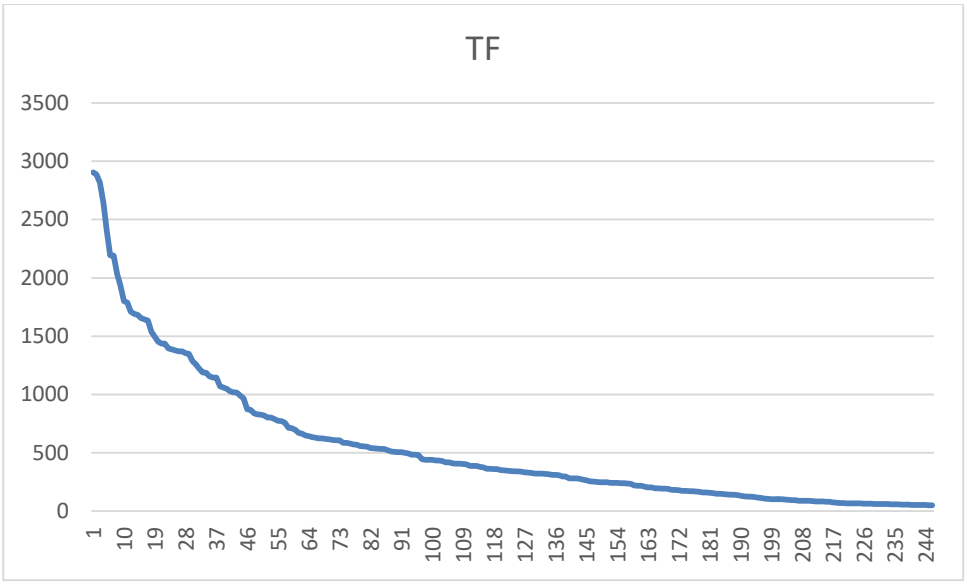
$$Knee = \begin{cases} ۰, & |N| < m \\ i, & |N| \geq m\ , \max(\ |av_i| < T\ ) \end{cases} \tag{۴-۳}$$

در رابطه فوق v_i میزان تغییرات تکرار کلمه نسبت به دو کلمه‌ی با فرکانس کمتر، $|av|$ میانگین تغییرات نسبت به m کلمه در اطراف خود است (میانگین گرفتن به خاطر نرم[۱] کردن تغییرات و مقدار m میزان این یکنواختی را تعیین می‌کند). برای مثال در شکل شکل ۳-۳ و شکل ۳-۴ نشان داده شده است که بعد از پالایش اولیه حدود ۵۰۰ کلیدواژه باقی‌مانده‌اند که بعد از پالایش ثانویه با استفاده از روابط فوق حدودا به ۲۴۴ کلمه کاهش یافته است.

[۱] Smoothing



شکل ۳-۳ فراوانی کلمات کاندید انتخاب‌شده رشته مکانیک



شکل ۳-٤فراوانی کلمات نهایی انتخاب‌شده رشته مکانیک

تا این مرحله با استفاده از روش‌های پالایش ذکر شده لیست کلیدواژهای پالایش شده و کلمات با اهمیت‌تر جدا شده‌اند. این کلمات، کلیدواژه‌هایی هستند که در تحلیل‌روند مورد استفاده قرار می‌گیرند. حال می‌بایست با استفاده از روشی کلمات را بازنمایی کرده و برای مدل‌سازی موضوعی آماده کرد.

قبل از موضوع بندی یک حوزه، می‌بایست حوزه‌های اسناد را از یکدیگر جدا کرد. برای تفکیک حوزه‌های مختلف و مقایسه روش ارائه‌شده می‌بایست مجموعه‌داده[1]های سند-کلمه تشکیل شوند. از آنجا که روش‌های مختلفی برای بازنمایی کلمات وجود دارند در این رساله برخی از متداول‌ترین و معروف‌ترین روش‌های بازنمایی پیاده شده و با روش پیشنهادی مقایسه شده‌اند. بنابراین مجموعه‌داده‌هایی که در این رساله به منظور تحلیل‌روند و پیش‌بینی فناوری استفاده شده‌اند به قرار زیر هستند:

- تشکیل مجموعه‌داده ماتریس سند-کلمه بر اساس TF (به‌منظور تفکیک اسناد حوزه‌های مختلف در ابتدای گام مدل‌سازی)؛
- تشکیل مجموعه‌داده ماتریس سند-کلمه بر اساس TF*IDF (به‌منظور تفکیک اسناد حوزه‌های مختلف در ابتدای گام مدل‌سازی)؛
- تشکیل مجموعه‌داده با استفاده از احتمال هم‌رخدادی دو کلمه بر اساس میزان تکرار آن‌ها (روش اصلی پیشنهادی در این رساله)؛

لازم به ذکر است که با استفاده از کلیدواژه‌های استخراج‌شده در مرحله قبل، ماتریس سند-کلمه را تشکیل می‌دهیم. این مجموعه‌داده‌ها با هدف بررسی و مقایسه عملکرد کلیدواژه‌های استخراج‌شده از مجموعه متون و بررسی عملکرد دسته‌بندها (یا خوشه‌بندها) در متون علمی فارسی و انتخاب بهترین دسته‌بندی (یا خوشه‌بندی) صورت پذیرفته است. علاوه بر آن با استفاده از این مجموعه‌داده‌ها اسناد حوزه‌های مختلف را از یکدیگر تفکیک خواهند شد تا در گام بعدی موضوع‌های هر حوزه استخراج شوند. مجموعه متون سند-کلمه به دو روش: تشکیل مجموعه‌داده ماتریس سند-کلمه بر اساس TF و بر اساس TF*IDF ساخته می‌شود. *سهم علمی در این بخش بررسی روش‌های دسته‌بندی و خوشه‌بندی در متون علمی فارسی و بررسی تاثیر روش استخراج کلیدواژه‌های ارائه شده در این رساله است.*

۳-۳-۱- روش پیشنهادی بازنمایی کلمات (تشکیل مجموعه‌داده با استفاده از

[1] Data Set (DS)

هم‌رخدادی کلمات)

یکی از مهمترین گام‌ها در پردازش زبان طبیعی بازنمایی کلمات است. از آنجا که کلمات موجود در یک متن از نظر معنایی با یکدیگر در ارتباط هستند (Harris۱۹۵۴)، یکی از روش‌های کشف موضوع، تشخیص، ترسیم نقشه علمی[۱] و تحلیل‌روند استفاده از هم‌رخدادی[۲] کلمات است. هم‌رخدادی کلمات می‌تواند ساختار فکری و زمینه تحقیقاتی یک حوزه را آشکار کند(Hu and Hu et al. (۲۰۱۵Zhang ۲۰۱۳).در ماتریس هم‌رخدادی هر سطر و یا ستون بیانگر میزان هم‌رخدادی یک کلمه با سایر کلمات است.

در این زیربخش روشی ارائه کردیم که با استفاده از آن میزان هم‌رخدادی کلمات بر اساس میزان احتمال رخداد دو کلمه محاسبه می‌شود. هدف از این بخش یافتن روشی است که بهترین اثر را در پیدا کردن و تفکیک موضوع‌های یک یا چند حوزه مختلف دارا باشد. روش ارائه‌شده در تفکیک حوزه‌ها (خوشه‌بندی حوزه‌ها با خلوص بالا) نیز از کارایی بالایی برخوردار است. در بخش‌های آتی با استفاده از روش‌های خوشه‌بندی نرم[۳] و سخت[۴] همراه با هم‌رخدادی کلمات عناوین مختلف را استخراج خواهیم کرد. روش‌های محاسبه میزان هم‌رخدادی کلمات به‌قرار زیر است:

- تعداد اسنادی که در آن‌ها دو کلمه موردنظر هم‌رخداد شده‌اند[۵].
- حداقل تعداد هم‌رخدادی‌های دو کلمه[۶]
- حداکثر تعدد هم‌رخدادی‌های دو کلمه[۷]
- احتمال هم‌رخدادی مشترک دو کلمه بر اساس میزان تکرار آن‌ها (روش‌پیشنهادی در این رساله)

---

[۱] Scientific Roadmap

[۲] Co-occurrence

[۳] Soft Clustering

[۴] Crisp

[۵] Document Frequency (DF)

[۶] Co-TF min

[۷] Co-TF max

- تشکیل مجموعه متون و مجموعه‌داده با استفاده از میزان شباهت کلمات بر اساس مدل بردار معنایی[1] و روش Vec۲Word.

در پژوهش‌های گذشته، هم‌رخدادی کلمات عموماً برحسب شمارش تعداد اسنادی که دو کلمه در آن رخداده‌اند محاسبه‌شده است. در این روش تعداد رخدادهای کلمات در متن در نظر گرفته نمی‌شوند و از آنجا که میزان رخداد کلمه در یک سند اهمیت کلمه را در آن سند نشان می‌دهد، لذا در این نوع از روش‌ها اهمیت کلمه در تعداد رخدادهای دو کلمه لحاظ نشده است. بنابراین بهتر است که این اهمیت را در محاسبه رخدادهای دو کلمه لحاظ کنیم. برای محاسبه هم‌رخدادی روش‌های دیگری می‌توان ارائه کرد که بر اساس تعداد رخداد دو کلمه باهم در یک متن محاسبه می‌شوند. روابط زیر این نحوه محاسبه این دو روش را نشان می‌دهد:

$$min\ CO - TF_{ij} = \sum_{d=1}^{D} \min\left(TF_i, TF_j\right)_d \qquad (۵-۳)$$

$$max\ CO - TF_{ij} = \sum_{d=1}^{D} \max\left(TF_i, TF_j\right)_d \qquad (۶-۳)$$

در روش پیشنهادی از احتمال هم‌رخدادی مشترک دو کلمه بر اساس میزان تکرار آن‌ها برای نمایش کلمات استفاده می‌شود. برای مؤثر نمودن میزان فراوانی کلمات، احتمال وقوع دو کلمه را به میزان احتمال وقوع هریک از دو کلمه در مقالات تقسیم کرده‌ایم (رابطه ۳-۷). این عمل باعث می‌شود تا کلماتی که هم‌رخدادی یکسانی دارند، لیکن اسناد کمتری تکرار شده‌اند نسبت به کلماتی که در اسناد بیشتری تکرار شده‌اند، امتیاز بیشتری داشته باشند و ضریب بالاتری را به خود اختصاص دهند. از طرف دیگر درروش پیشنهادی در محاسبه میزان هم‌رخدادی دو کلمه، تعداد تکرار هر یک از آن دو کلمه، به‌صورت مجزا، در یک سند نیز لحاظ شده است که این امر باعث مؤثر شدن میزان اهمیت یک کلمه در یک سند می‌شود. درواقع با این کار هم‌رخدادی دو کلمه، در حالتی که دو

[1] Vector semantics

کلمه بیش از یک‌بار در یک سند تکرار شده باشند نسبت به حالتی که تنها یک‌بار در سند آمده باشند، بیشتر خواهد بود.

بنابراین اگر رویداد مربوط به رخداد کلیدواژه iام را k_i و رویداد مربوط به رخداد کلیدواژه j ام را k_j در نظر بگیریم، آنگاه درایه ماتریس هم‌رخدادی C_{ij} که معادل احتمال رخداد این دو کلیدواژه باهم است، به‌صورت زیر محاسبه می‌شود:

$$C_{ij} = \frac{P(k_i \cap k_j)}{P(k_i \cup k_j)} \qquad (۳-۷)$$

که در آن $P(k_i \cup k_j)$ احتمال رخداد هریک از کلیدواژه‌ها در اسناد موردبررسی و $P(k_i \cap k_j)$ احتمال رخداد دو کلیدواژه k_i و k_j باهم در یک سند است. احتمال رخداد دو واژه باهم در یک سند، بر اساس تعریف احتمال شرطی به‌صورت زیر است:

$$P(k_i \cap k_j) = P(k_i|k_j) * P(k_j) \qquad (۳-۷)$$

احتمال شرطی $P(k_i|k_j)$ به این مفهوم است که با فرض اینکه k_j در یک سند مشاهده‌شده، احتمال اینکه k_i هم در همان سند مشاهده شود، چقدر است. بنابراین اگر D_i مجموعه کلیه اسنادی باشد که k_j را در بر دارد، آنگاه این احتمال شرطی را به‌صورت زیر محاسبه می‌کنیم:

$$P(k_i|k_j) = \frac{\sum_{d \in D_i} \min\left(freq(k_i, d), freq(k_j, d)\right)}{\sum_{d \in D_i}\left(freq(k_j, d)\right)} \qquad (۳-۸)$$

منظور از $freq(k_j, d)$ میزان تکرار کلمه j ام در سند d است. برای محاسبه احتمال رخداد یک کلمه به‌شرط رخ دادن کلمه دیگر (یعنی رابطه (۳-۸)) در یک سند، از تقسیم حداقل تعداد تکرار هر دو کلمه بر روی کل کلمات در آن سند استفاده‌شده است. برای مثال اگر کلمه اول، سه بار در سند آمده باشد و کلمه دوم ۴ بار در همان سند آمده باشد و تعداد کل کلمات آن سند ۱۰۰ کلمه باشند نتیجه رابطه (۳-۸) برابر با $\frac{۳}{۱۰۰}$ خواهد بود. این بدان علت است که تعداد هم‌رخدادی دو کلمه هم‌زمان نباید از تکرار هر یک از آن‌ها بیشتر باشد بنابراین در رابطه (۳-۸) حداقل تکرار دو کلمه

لحاظ شده است. احتمال رخداد یک کلمه یعنی $P(k_j)$ از رابطه (۳-۹) محاسبه می‌شود که در آن مخرج کسر نشان‌دهنده تعداد کلمات سند است:

$$P(k_j) = \frac{\sum_{i \in D} freq(k_i, d)}{\sum_{j \in Corpus} |D_j|} \quad (۳-۹)$$

از حاصل‌ضرب رابطه (۳-۸) و (۳-۹) می‌توان به این نتیجه رسید که احتمال هم‌رخدادی دو کلمه j و i به‌صورت زیر است:

$$P(k_i \cap k_j) = \frac{\sum_{i \in D} \min\left(freq(k_i, d), freq(k_j, d)\right)}{\sum_{j \in Corpus} |D_j|} \quad (۳-۱۰)$$

ازآنجاکه مخرج کسر وابسته به کلمات k_j و k_i نیست و برای تمام کسرها یک مقدار دارد می‌توان برای محاسبه احتمال اشتراک دوکلمه‌ای از آن صرف‌نظر کرد؛ بنابراین این احتمال تنها متناسب با صورت کسر به‌صورت رابطه (۵) در نظر گرفته می‌شود:

$$P(k_i \cap k_j) \propto \sum_{d \in D} \min\left(freq(k_i, d), freq(k_j, d)\right) \quad (۳-۱۱)$$

از طرفی دیگر برای به دست آوردن $P(A \cup B)$ از رابطه (۳-۱۲) استفاده خواهیم کرد:

$$P(k_i \cup k_j) = P(k_i) + P(k_j) - P(k_i \cap k_j)$$
$$= \frac{\sum_{i \in D} freq(k_i, d) + \sum_{i \in D} freq(k_j, d) - \sum_{d \in D_i} \min\left(freq(k_i, d), freq(k_j, d)\right)}{\sum_{d \in D_i} \left(freq(k_j, d)\right)} \quad (۳-۱۲)$$

بنابراین هر درایه ماتریس هم‌رخدادی، یعنی هم‌رخدادی دو کلمه i و j، به‌صورت زیر محاسبه می‌شود:

$$C_{ij} = \frac{\sum_{i \in D} \min\left(freq(A_i, B_i)\right)}{\sum_{i \in D} freq(k_i|d) + \sum_{i \in D} freq(k_{ij}, d) - \sum_{d \in D_i} \min\left(freq(k_i, d), freq(k_j, d)\right)} \quad (۳-۱۳)$$

علاوه بر روش پیشنهادی فوق برای بازنمایی کلمات و کلیدواژه‌ها، روش‌های معنایی دیگری نیز وجود دارد که در بخش پیشینه پژوهش نیز عنوان شده است. یکی از پرکاربردترین آنها که در سال‌های اخیر نیز مورد توجه قرار گرفته است، روش کلمه-به-بردار (vec۲word) است. این روش

در تشخیص و استخراج روابط بین کلمات کاربرد فراوان می‌تواند داشته باشد. در این رساله نیز برای بررسی‌های بیشتر از این روش برای تحلیل‌ها و مقایسات بیشتر استفاده کرده‌ایم که نحوه بکارگیری آن در بخش‌های بعدی توضیح داده می‌شود.

۳-۴- استخراج موضوع‌های حوزه‌های علمی

همان‌طور که در فصول قبل بیان شد به دلیل رشد و توسعه حوزه‌های علمی در طول زمان، موضوع‌های متنوعی در دل این حوزه‌ها شکل‌گرفته و حتی برخی از موضوع‌ها نیز از بین رفته‌اند. به دلیل حجم بالای اسناد و نیز با توجه به وجود روش‌های ماشینی برای پردازش آنها، رویکردهای متنوعی برای کشف موضوع‌های جدید در تحقیقات گذشته ارائه‌شده است (برای مثال مراجعه کنید به (Leydesdorff and Nerghes ۲۰۱۷, Cheng et al. ۲۰۱۴)). علاوه این حتی در صورت داشتن منابع انسانی برای بررسی اسناد توسط نیروی متخصص، اشتباهات انسانی و پیش‌زمینه‌های فکری فرد متخصص می‌تواند چالش‌هایی را در استخراج عناوین ایجاد نماید. یکی از اهداف این رساله کشف *خودکار* حوزه‌های علمی است تا بتوانیم روند آن‌ها را موردبررسی و تحلیل قرار دهیم.

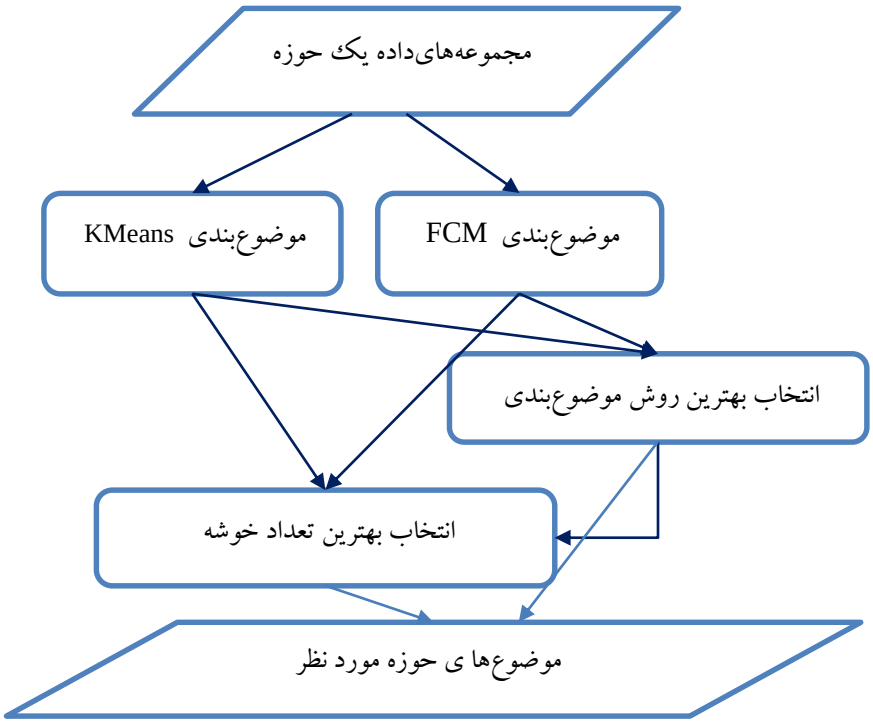
استفاده از روابط معنایی موجود در اصطلاح‌نامه یکی دیگر روش‌های خوشه‌بندی است که می‌تواند در خوشه‌بندی کمک شایانی انجام دهد. بررسی که توسط کریمی در سال ۲۰۱۸ انجام شد نشان داد که روابط موجود در اصطلاح‌نامه عمومی با کلمات موجود در چکیده سازگاری ندارند که این می‌تواند بخاطر ارتباط علمی بین کلمات در چکیده‌های پارسا باشد. بنابراین برای بهبود در خوشه‌بندی و تشخیص ارتباط بین کلمات نمی‌توان از اصطلاح‌نامه استفاده کرد.

انواع روش‌ها کشف موضوع‌ها را در فصل‌های گذشته بررسی کرده‌ایم. به‌طور خلاصه می‌توان روش‌های کشف موضوع را به‌صورت استفاده از خوشه‌بندی‌های کلمات و استفاده از روش‌های احتمالاتی (مانند LDA[۱]) در نظر گرفت. در خوشه‌بندی کلماتی که از نظر معنایی مشابه هم هستند در یک خوشه قرار می‌گیرند و یک گروهی از کلمات را تشکیل می‌دهند (Wang et al. ۲۰۱۵) که این گروهی کلمات را موضوع می‌توان در نظر گرفت (de la Hoz-Correa, Muñoz-Leiva,

[۱] Linear Discriminant Analysis

۲۰۱۸and Bakucz) و می‌توان با تحلیل خوشه‌ها و هم‌رخدادی زمینه‌های تحقیقاتی را تحلیل کرد (۲۰۱۵Hu and Zhang).

در این بخش از رساله با استفاده از مجموعه‌داده‌های پیشنهادی ، روشی برای موضوع‌بندی اسناد ارائه شده‌است. رویکرد پیشنهادی در این بخش به شرح شکل ۳-۵ است. این شکل مراحل استخراج موضوع‌ها را از مجموعه‌داده‌ها نشان می‌دهد که با استفاده از روش‌های کا-میانه، خوشه‌بندی فازی عنوان‌های موضوع‌های اسناد کشف می‌شود. با مقایسه روش‌ها بهترین روش برای انتخاب موضوع‌های انتخاب خواهد شد. بعد از انتخاب روش استخراج موضوع بهترین حالت ازنظر تعداد موضوع‌ها با استفاده از معیارهای ارزیابی درونی انتخاب خواهد شد.



شکل ۳-۵ نمودار چگونگی کشف خودکار موضوع

۳-۴-۱- موضوع‌بندی با استفاده از الگوریتم‌های خوشه‌بندی کا-میانه و فازی

در مرحله اول با استفاده از الگوریتم خوشه‌بندی کامیانه و فازی بر اساس روابط بین هم‌رخدادی کلمات کلیدی استخراج‌شده موضوع‌هایی را استخراج می‌کنیم. هر دو روش خوشه‌بندی روش‌های مبتنی بر تقسیم‌بندی داده‌ها هستند که برای حجم زیاد داده مناسب هستند (Zhao, Li, and Cang ۲۰۱۵). برای بررسی میزان دقت این موضوع‌ها از دو روش استفاده خواهیم کرد. در روش اول بر

اساس پارامترهای درونی خوشه‌ها را مورد ارزیابی قرار می‌دهیم و حالتی بهتر است که پارامترهای درونی خوشه بهتری داشته باشد. در روشی دیگر با استفاده از این موضوع‌ها، سندهایی که از آن‌ها این موضوع‌ها استخراج‌شده‌اند را خوشه‌بندی کرده و میزان خلوص سندها را مورد تحلیل قرار خواهیم داد.

۳-۴-۲- کشف موضوع با استفاده از الگوریتم LDA

یکی از معروف‌ترین روش‌های مدل‌سازی موضوعی استفاده از روش LDA است. برای جامعیت و مقایسه کار از این روش نیز برای کشف موضوع استفاده خواهیم کرد. برای مقایسه این روش با روش پیشنهادی در این رساله، با استفاده از موضوع‌های کشف‌شده توسط روش LDA سندها را دسته‌بندی خواهیم کرد و میزان درستی و خلوص سندها را با روش پیشنهادی مقایسه خواهیم نمود.

۳-۴-۳- تعیین تعداد موضوع‌ها (خوشه‌ها)

از آنجا که کلیدواژه‌ها نماینده‌ای از یک موضوع هستند (Karami et al. ۲۰۱۵, He ۱۹۹۹)، با استخراج کلیدواژه‌های مجموعه‌ای از اسناد علمی یک حوزه می‌توانیم موضوع‌های مرتبط با آن حوزه را پیدا کنیم. بسیاری از این کلمات در یک راستا هستند و می‌توانند یک موضوع را نشان‌دهنده. بنابراین می‌توانیم با دسته‌بندی کلماتی[۱] که باهم به‌کاربرده می‌شوند فضای حالت موجود را کاهش دهیم و به موضوع‌های یکتایی برسیم. ازآنجاکه تعداد موضوع‌های دقیق یک حوزه علمی که در اسناد استخراج‌شده مشخص نیستند[۲]، نیاز به روش خاصی است که بتوانیم بهترین حالت از تعداد موضوع‌های موجود را پیدا کنیم. پیدا کردن مقدار بهینه این تعداد در مجموعه اسناد می‌تواند بر اساس ویژگی ذاتی موجود در آن اسناد صورت پذیرد.

---

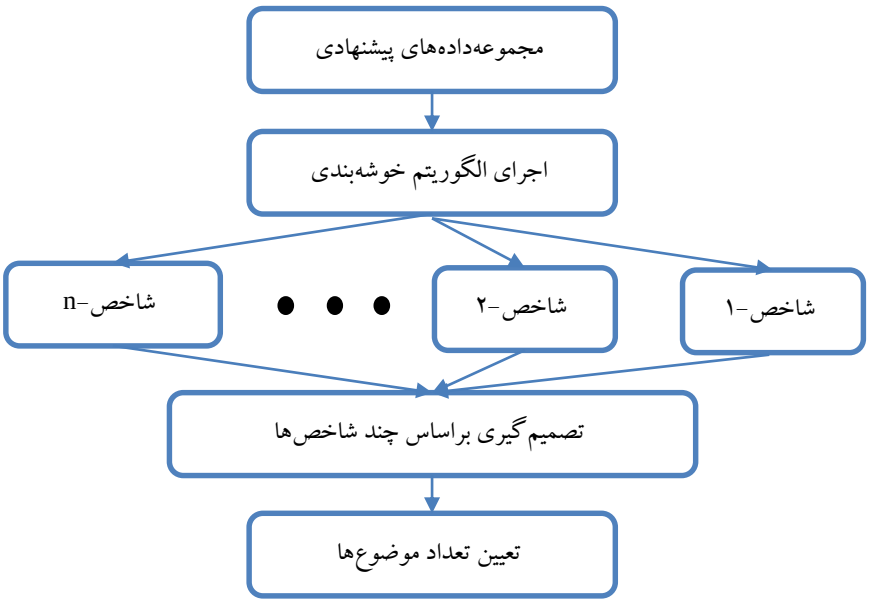
[۱] توسط الگوریتم‌های خوشه بندی

[۲] تعداد خوشه‌های هدف (و فاصله بین خوشه ای) از پیش معلوم نیست و می‌تواند با توجه به متن تعداد خوشه‌های متفاوتی داشته باشیم. همین امر باعث افزایش پیچیدگی الگوریتم و کاهش دقت آن می‌شود.

در این رساله ما ویژگی‌های اسناد علمی را احتمال وقوع مشترک کلیدواژه‌ها در مجموعه چکیده‌های اسناد علمی تعریف کرده‌ایم. نحوه محاسبه این ویژگی در بخش قبل آمده است. سپس با استفاده از این ویژگی‌ها موضوع‌ها را کشف کرده و این موضوع‌ها را توسط شاخص‌های ارزیابی موردبررسی قرار خواهیم داد.

روش‌های متفاوتی برای پیدا کردن تعداد خوشه‌ها (موضوع‌ها) ارائه‌شده است که در فصل پیشینه پژوهش به برخی از آن‌ها پرداخته‌شده است. تعداد خوشه‌ای بهتر است که نتایج ارزیابی خوشه‌بندی بهتری را ارائه دهد. برای ارزیابی خوشه‌بندی روش‌های متعددی وجود دارد که در کل می‌توان آن‌ها را به دودسته ارزیابی‌های درونی (ارزیابی‌هایی که بر اساس پارامترها ذاتی داده‌ها و به‌صورت غیر نظارت‌شده صورت می‌پذیرد) و ارزیابی‌های بیرونی (پارامترهای ارزیابی‌های نظارت‌شده صورت می‌پذیرد) تقسیم‌بندی کرد. ازآنجاکه در مسئله ما برچسبی برای موضوع‌های موجود در یک حوزه علمی وجود ندارد، روش‌های ارزیابی بیرونی کارایی نداشته و می‌بایست با استفاده از روش‌های ارزیابی درونی (غیر نظارت‌شده) به کشف تعداد بهینه‌ای از موضوعات پی‌ببریم.

چهارچوب ارائه‌شده برای پیدا کردن تعداد موضوع‌ها به این صورت است که بعد از خوشه‌بندی با تعداد خوشه‌های متفاوت و محاسبه شاخص‌های ارزیابی، با کمک این شاخص‌های ارزیابی داخلی بهترین حالت خوشه‌بندی پیشنهاد داده خواهد شد. از این روش در (Peng, Zhang, Kou, and



(Shi۲۰۱۲) و(Peng, Zhang, Kou, Li, et al. ۲۰۱۲) برای ارزیابی تعداد خوشه‌ها استفاده‌شده است. شکل ۳-۷ این چهارچوب را نشان خواهد داد. ما در این پیاده‌سازی از معیارهای درونی سیلوهت[۱] (Rousseeuw ۱۹۸۷) و میزان فاصله از مرکز خوشه‌ها (SSE) استفاده خواهیم کرد. لازم به یادآوری است که درصورتی‌که داده‌ها از ترکیب دو حوزه متفاوت برچسب دار شده تشکیل‌شده باشند، با استفاده از برچسب‌های موجود، می‌توان از شاخص ارزیابی بیرونی (مانند شاخص خلوص) استفاده کرد تا میزان موضوع‌های ترکیب دو حوزه متفاوت را پیداکرده وپیشنهاد داده شوند

۳-۵- تحلیل‌روند موضوع‌های کشف‌شده

تحلیل‌روند به منظور کسب اطلاعات با اهمیت در بازه‌های زمانی متفاوت برای رسیدن به الگوهایی جهت تصمیم‌گیری است. تحلیل‌روند موضوع‌های کشف شده از هم‌رخدادی کلمات می‌تواند در پیش‌بینی آینده مورد استفاده قرار گیرد (Callon, Courtial, and Laville ۱۹۹۱). تحلیل‌روند به دو صورت تحلیل بین‌موضوعی (میزان میان‌رشته‌ای بودن و مرکزیت موضوع) و تحلیل درون‌موضوعی (میزان بلوغ و روند تکامل موضوع) انجام می‌شود.

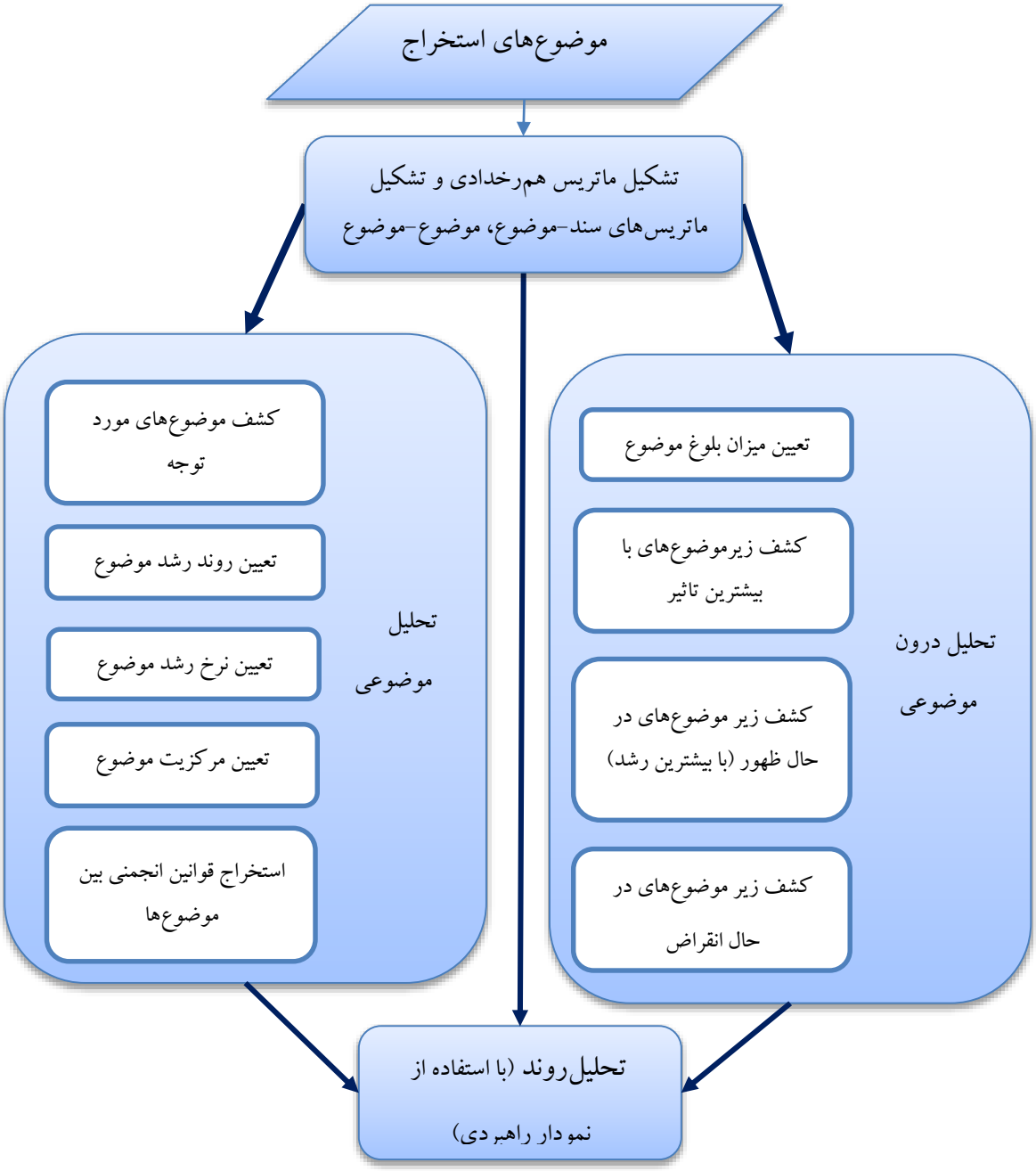
در این بخش علاوه بر ارائه دو شاخص برای محاسبه میزان بلوغ و مرکزیت موضوع، روشی متفاوت برای تحلیل‌روند موضوع‌ها با استفاده از نمودار راهبردی ارائه کرده‌ایم. در روش تحلیل پیشنهادی ترسیم نمودار راهبردی با استفاده از ماتریس هم‌رخدادی موضوعی و ماتریس سند-موضوع صورت می‌گیرد. از طرفی برای تحلیل‌های بهتر با استفاده از قوانین انجمنی روابط چندگانه بین موضوع‌ها استخراج خواهد شد. که بر اساس پیشینه‌پژوهش استفاده از نوع تحلیل برای تحلیل‌روند یافت نشده است. از آنجا که هدف از نمودار راهبردی بررسی وضعیت و روند موضوعات است، ویژگی بارز روش تحلیل ارائه‌شده، تحلیل کلان‌تر موضوعات و حذف نویز موضوع‌ها (به کلماتی نویز می‌گوییم که نشان‌دهنده موضوع اصلی اسناد نیستند ولی بخاطر وجود در چکیده اسناد در محاسبات لحاظ می‌شوند) است.

[۱] Silhouette

برای ترسیم این نوع نمودار راهبردی با استفاده از موضوع‌های تشکیل شده، میزان مشارکت هر موضوع در هر سند علمی محاسبه می‌شود. این محاسبه بر حسب میزان استفاده سند از کلمات هر موضوع (وزن و تکرار کلمات داخل چکیده خود است) انجام می‌شود. بعد از اینکه میزان مشارکت هر موضوع در هر سند مشخص شد، می‌بایست موضوع‌های اصلی هر سند را تشخیص دهیم. از آنجا که اسناد از کلمات مختلفی تشکیل شده‌اند، با استفاده از آستانه گذاری موضوع‌هایی که مشارکت بیشتری دارند را به آن سند نسبت می‌دهیم (موضوع‌هایی که بیشتر از ۱۰ درصد از کل کلمات آن سند را دارا باشند). این آستانه گذاری برحسب نظر متخصص انجام خواهد شد.

در گام بعدی مجموعه‌داده (ماتریس) سند-موضوع تشکیل خواهیم داد. این ماتریس اسناد را بر حسب موضوع‌ها بیان می‌کند و مشخص می‌کند که هر سند از چه موضوع‌هایی تشکیل شده است. با استفاده از این ماتریس ماتریس هم‌رخدادی بین موضوع‌ها را محاسبه خواهیم کرد که ماتریس جدید موضوع-موضوع تشکیل خواهد شد. با استفاده از ماتریس سند-موضوع، میزان استفاده از هر موضوع در مجموعه اسناد به عنوان بلوغ (شاخص چگالی) محاسبه می‌شود و با استفاده از ماتریس موضوع-موضوع میزان ارتباط بین موضوع‌ها (شاخص مرکزیت) محاسبه خواهد شد.

در تحلیل درون موضوعی نیز روند رشد کلیدواژه‌ها (را که می‌توان زیر موضوع‌ها دانست) مورد تحلیل قرار خواهند گرفت. شکل ۳-۷ مراحل تحلیل‌روند پیشنهادی در رساله را نشان می‌دهد. در هر نوع تحلیل شاخص‌ها و نمودارهای مخصوص به آن تحلیل محاسبه و ترسیم می‌شود. برای بصری سازی، علاوه بر نمودار راهبردی از نمودارهای فراوانی برحسب زمان نیز استفاده می‌شود.



شکل ۳-۷ مراحل تحلیل‌روند پیشنهادی در رساله

۳-۵-۱- تحلیل موضوعی (خوشه‌ای)

بعد از استخراج موضوع‌ها، با استفاده از ماتریس‌های هم‌رخدادی کلمات و هم‌موضوعی، ارتباط بین موضوع‌های مختلف بر اساس ارتباط بین کلمات دو موضوع در ماتریس هم‌رخدادی و ارتباط موضوع‌ها در ماتریس هم‌موضوعی به دست آورده می‌شوند. با استفاده از این ارتباط به میزان

میان‌رشته‌ای بودن موضوع‌ها و روند رشد آن‌ها می‌توان پی برد. از طرفی با استفاده از میزان تکرار موضوع‌ها در اسناد و یا میزان تکرار کلمات در هر موضوع میزان بلوغ آن حوزه محاسبه خواهد شد. بنابراین می‌توان از منظرهای زیر، موضوع‌ها را بررسی نمود:

- **شاخص بلوغ یک موضوع** (TD[1]): با بررسی روند رشد این شاخص، می‌توان به میزان محبوبیت موضوع در بین محققین و پژوهشگران پی‌برد. رشد موضوع‌ها بر اساس شاخصی ترکیبی از میزان تکرار هم‌رخدادی کلمات بین حوزه‌های مختلف و میزان ارتباط آن‌ها موردبررسی قرار می‌گیرد. از طرفی دیگر می‌توان ارتباط بین‌کلمات را با استفاده از این شاخص بررسی کرد. هرچه این مقدار بیشتر باشد خوشه بیشتر به بلوغ و تکامل رسیده است (An and Wu ۲۰۱۱, Shen, Li, and Gu ۲۰۱۳). برخی روش‌های محاسبه این شاخص در فصل دوم بیان‌شده است.

شاخص اول ارائه‌شده در این رساله برای محاسبه بلوغ یک موضوع نرمال شده حاصل جمع وزن اتصالات درون موضوع و تکرار کلمات موضوع (که نشان‌دهنده اهمیت کلمه است (Hu and Zhang ۲۰۱۵)) بر حسب تعداد کلمات درون خوشه است. شاخص دوم جمع تعداد اسنادی که موضوع مورد نظر را در بر می‌گیرد. این دو شاخص با استفاده از روابط ۳-۱۴ و ۳-۱۵ محاسبه خواهند شد.

$$TD_i = \frac{\sum_{i,j \in Cluster_k} WeightOf\,\mathrm{Link_{ij}} + \sum_{i \in Cluster_k} Frequency_i}{\sum MemberOfCluster_k} \qquad (۱۴-۳)$$

$$TD_i = \left|\{d_j | T_i \in d_j, d_j \in D\}\right| \qquad (۱۵-۳)$$

که در این رابطه d_j اسنادی هستند که در مجموعه اسناد استخراج‌شده D قرار دارند. T_i موضوع iام را نشان می‌دهد. TD تعداد اسنادی که موضوع موردنظر رادارند است.

[1] Topic Density

- **نرخ رشد موضوع** (TGR[۱] Index): با استفاده از نرخ رشد، میزان افزایش محبوبیت یک حوزه نسبت به سال گذشته نشان داده می‌شود. با این شاخص می‌توان مشخص کرد که در طول سال‌های متوالی چه تغییراتی بر روی آن موضوع صورت پذیرفته است. شاخص نرخ رشد موضوع از تفاضل میزان فعالیت انجام‌شده در سال گذشته و سال جاری محاسبه می‌شود.

$$TGR_i = TD_i - TD_{i-۱} \qquad (۱۶-۳)$$

- **شاخص مرکزیت موضوع** (TC[۲]): شاخص مرکزیت در نمودار راهبردی بر اساس رابطه‌ی یک موضوع با دیگر موضوع‌ها محاسبه می‌شود. این محاسبه میواد از طریق هم‌رخدادی کلیدواژه‌ها یا ماتریس هم‌رخدادی موضوع‌ها محاسبه شود. از این روش می‌توان در تشخیص موضوع‌های میان‌رشته‌ای نیز استفاده کرد. به برخی از روش‌های محاسبه این شاخص در فصل دوم اشاره شده است.

در این رساله روش استخراج روابط بین‌موضوع‌ها با استفاده از ماتریس سند-موضوع و قوائد انجمنی ارائه‌شده و شده است. علاوه بر آن شاخص مرکزیت موضوع با استفاده از میزان ارتباط موضوع‌ها در ماتریس هم‌رخدادی موضوعی محاسبه شده است. شاخص دیگری در سطح کلمات پیشنهاد شده، که میزان ارتباط بین کلمات دو خوشه را براساس احتمال هم‌رخدادی دو کلمه که در بخش ۳-۳-۱- ارائه گردید محاسبه می‌شود. روابط ۳-۱۷ و ۳-۱۸ محاسبه مرکزیت را براساس روش‌های پیشنهادی نشان می‌دهند.

$$TC_i = \frac{\sum_{\substack{i \in Cluster_k, \\ j \in Cluster_m}} WeightOf\,\mathrm{Link_{ij}}}{\sum NOTMemberOfCluster_k} \qquad (۱۷-۳)$$

[۱] Topic Growth Rate

[۲] Topic Centrality

$$TC_{i_j} = \left|\{d_k | \, T_j \in d_k \text{ و } T_i \in d_k, d_k \in D\}\right| \qquad (۱۸-۳)$$

رابطه فوق تعداد اسنادی که دو موضوع را باهم دارند نشان می‌دهد.

- **تحلیل موضوعی با استفاده از قوانین انجمنی:** سهم علمی دیگر این رساله پیشنهاد بررسی ارتباط بین موضوع‌ها و احتمال هم‌رخدادی با استفاده از قوانین انجمنی است. بعد از تشکیل موضوع‌ها، مشخص خواهیم کرد که هر سند متعلق به چه موضوع یا موضوع‌هایی است. بنابراین می‌توان با استفاده از موضوع‌های هر سند ماتریس جدید سند-موضوع را تشکیل داد. با استفاده از این ماتریس سند-موضوع و قوانین انجمنی می‌توانیم روابط موجود در بین موضوع‌ها را استخراج کرد. استفاده از قوانین انجمنی ازآنجا قابل اهمیت است که به خاطر کثرت تعداد پارساها در مجموعه متون و نیز ترکیب‌های مختلف موضوع‌ها استخراج این قوانین به‌صورت دستی توسط فرد متخصص امری بسیار زمان‌بر است.

۳-۵-۲- تحلیل درون موضوع

هرچند که کلمات در هر خوشه ارتباط معنایی با یکدیگر دارند، اما تحلیل کلیدواژه‌ها در هر موضوع نیز دارای معنی است. تحلیل‌های زیر برای کلیدواژه‌های درون یک موضوع در نظر گرفته شده است:

- کلماتی که بیشترین میزان تأثیر **(بیشترین فراوانی هم‌رخدادی)** را در فراوانی خوشه دارند. این کلمات می‌توانند زیرموضوع اصلی یک موضوع را نشان دهند.
- کلماتی که **بیشترین رشد** را در موضوع دارند. این کلمات زیر موضوعی که در آن حوزه موردتوجه قرار می‌گیرد را نشان می‌دهند یا به‌اصطلاح حوزه‌های در حال ظهور را نشان می‌دهند.
- کلماتی که بیشترین **نرخ کاهش فراوانی** رادارند را به‌منظور تشخیص زیرموضوع‌های قدیمی و در حال انقراض استفاده کرد.

۳-۶- جمع‌بندی

تحلیل‌روند یکی از روش‌های پیش‌بینی است که با استفاده از بررسی داده‌های پیشین انجام می‌شود. روش‌های گوناگونی برای تحلیل‌روند ارائه‌شده است. تحلیل‌روند خودکار در چهار گام اصلی انجام شده است: بررسی و شناخت داده‌ها، تحلیل‌خودکار متون و استخراج خودکار کلیدواژه‌ها، بازشناسی الگو و مدل‌سازی موضوعی و تحلیل و ارزیابی روند (با نمودارها) است.

در این رساله برای نخستین بار روشی برای تحلیل‌روند با استفاده از یادگیری ماشین و به صورت تمام خودکار برای متون علمی فارسی ارائه‌شده است. علاوه بر آن در برخی از گام‌های تحلیل‌روند نوآوری‌هایی ارائه داده‌شده تا هر گام از تحلیل‌روند به صورت بهینه‌تر عمل کند. در انتها نیز یک روشی در خصوص تحلیل‌روند از دیدگاه کلانتر به موضوع‌ها ارائه‌شده است. روش پیشنهادی از بخش‌های مختلفی تشکیل شده است. هر بخش به‌صورت جداگانه موردبررسی قرار گرفت و نوآوری هر بخش به‌تفصیل در همان بخش مربوطه بیان گردید. در فصل بعد روش‌های ارائه‌شده پیاده‌سازی و موردبحث و بررسی قرار خواهند گرفت.

در گام نخست به شناخت خصوصیات اسناد در دسترس پرداخته شده است. روابط بین انواع کلیدواژه‌ها استخراج شده و توزیع آن‌ها به منظور استخراج خودکار کلیدواژه‌ها در عنوان و چکیده مورد بررسی قرار گرفته شده است. برای این منظور اسناد مورد نیاز استخراج و ساختار چکیده‌ها و کلیدواژه‌های پارسا را مورد بررسی و ارزیابی داده‌ایم. سپس اسناد را برای پردازش‌های زبان طبیعی باید به شکل مناسب و قابل پردازش برای سیستم‌های دیجیتال تبدیل کنیم. بنابراین در این بخش روشی برای استخراج کلیدواژه‌های اسناد ارائه کرده‌ایم. سپس با استفاده از روش پیشنهادی برای بازنمایی کلمات، کلیدواژه‌ها را به شکلی موثر برای استخراج موضوع‌های اسناد آماده ساخته‌ایم. در گام بعدی با استخراج الگوهای رفتاری بین کلمات (با استفاده از الگوریتم‌های کا-میانه و خوشه‌بندی فازی) موضوع‌های مخفی موجود در متون را مشخص خواهیم کرد. در ادامه شاخص‌های تحلیل‌روند را برای موضوع‌های استخراج شده محاسبه کرده و با استفاده از نمودار راهبردی موضوعات را بصری‌سازی و روابط بین موضوع‌ها را با استفاده از روش پیشنهادی قوانین انجمنی استخراج کرده و روند رشد موضوع‌ها مورد بررسی و تحلیل قرار خواهیم داد.

در این رساله چند رویکرد جدید برای تحلیل‌روند با استفاده از نمودار راهبردی و قوانین انجمنی نیز پیشنهاد شده است. علاوه‌بر محاسبه شاخص‌ها با استفاده از روش‌های متداول، شاخص تکمیلی که نقطه ضعف دخیل نکردن میزان تکرار کلمه را برطرف کرده است ارائه شده است. علاوه بر آن در رویکرد پیشنهادی محاسبه شاخص‌ها و ترسیم نمودار راهبردی بر اساس ماتریس هم‌رخدادی موضوعی محاسبه و نمایش داده شده‌اند. روش پیشنهادی هم‌رخدادی موضوعی باعث حذف کلمات نویز (کلماتی موجود در سند علمی که موضوع اصلی را بیان نمی‌کنند) تحلیلی با دید کلان تر (تحلیل از سطح کلمات علمی به سطح اسناد علمی) شده است.

روشی دیگری که برای تحلیل‌روند در این رساله ارائه شده است، استفاده از قوانین انجمنی برای استخراج روابط بین موضوع‌ها و پیش‌بینی استفاده از موضوع‌ها است. علاوه‌بر استفاده از نمودار راهبردی برای بصری‌سازی و تحلیل، روابط بین موضوع‌ها را با استفاده از روش پیشنهادی قوانین انجمنی استخراج شده‌اند. با استفاده از استخراج این روابط می‌توان وابستگی چندگانه موضوع‌ها به یکدیگر را ملاحظه کرد. بر اساس این روابط احتمال استفاده (میزان اطمینان بدست آمده از قواعد انجمنی) هر یک از موضوع‌ها توسط یک پژوهشگر مشخص می‌شود و می‌توان پژوهش‌های آتی افراد را پیش‌بینی کرد.

فصل ٤- پیاده‌سازی و ارزیابی

در این بخش روش پیشنهادی تحلیل‌روند را پیاده‌سازی خواهیم کرد. هر گام از روش‌پیشنهادی را بعد از پیاده سازی مورد بررسی و ارزیابی قرار خواهیم داد و با روش‌های شناخته شده گذشته مور مقایسه قرار خواهد گرفت.

۴-۱- تحلیل ساختار فراداده[۱]های پارسا

همانطور که به تفصیل در فصل قبل بیان شد می‌بایست در ابتدای داده کاوی ساختار داده‌های مورد استفاده را شناخت تا با استفاده از تحلیل‌ساختار داده از ویژگی‌های آن‌ها استفاده کرد. در این بخش با تحلیل‌ساختار فراداده‌های پارسا ی در دسترس درباره نحوه انتخاب کلیدواژه‌ها تصمیم‌گیری خواهیم کرد. به طور کلی در این بخش به پرسش‌های مطرح شده در بخش ۳-۲ پاسخ داده خواهد شد.

در بخش اول از گام و هدف پژوهش به بررسی ساختار چکیده پارسا خواهیم پرداخت. برای انجام این تحقیق، از پارساهای فارسی بین سال‌های ۱۳۹۰ تا ۱۳۹۳ که در سامانه گنج ایرانداک ثبت‌شده‌اند، استفاده‌شده است. یکی از معیارهای انتخاب پارساها وجود کلیدواژه‌های نویسنده پارسا و نمایه‌ساز حرفه‌ای بوده است. ازآنجاکه پارسا رشته‌های مهندسی در بین این رکوردها بیشتر بودند از بین پارساهای بازیابی شده تنها پارساهای رشته‌ی مهندسی انتخاب‌شده است.

تعداد این پارساها ۷۴۲ عدد هستند. پارساهای انتخاب‌شده می‌بایست شامل مواردی همچون چکیده، کلیدواژه‌های نویسنده، و توصیفگرها باشند. ازآنجاکه پارساهای رشته‌های مهندسی در این مجموعه ازنظر تعداد با یکدیگر یکسان نبودند، پارساهای گرایش‌های مهندسی که تعداد بیشتری

[۱] Meta-data

بودند انتخاب‌شده‌اند. پارساهای گرایش‌های مهندسی انتخاب‌شده شامل ۵۲۷ پارسا از رشته‌های مهندسی برق، مهندسی مکانیک و مهندسی عمران هستند.

جدول ٤-١ تعداد پارسای انتخاب شده در هر رشته و دانشگاه

	رشته‌های مهندسی	مهندسی برق	مهندسی مکانیک	مهندسی عمران	سه رشته
تعداد پارسا	۷۴۲	۲۰۴	۱۳۶	۱۸۷	۵۲۷
تعداد دانشگاه	۶۷	۲۸	۳۳	۴۱	۵۰

در بخش اول رابطه و الگوی کلیدواژه‌های انتخابی نویسنده و نمایه‌ساز استخراج شده محاسبه خواهد شد. نتایج به‌دست‌آمده نشان داده است که به‌طور متوسط برای هر پارسا ۳/۲ کلیدواژه توسط نویسنده و ۷/۳ توصیفگر توسط نمایه‌ساز انتخاب‌شده است. آمار تعداد کلیدواژه‌ها و توصیفگرها در جدول ۴-۲ آورده شده است. این جدول نشان می‌دهد که نمایه های نویسنده و نمایه‌ساز با هم متفاوت بوده و میزان اشتراک آن‌ها تنها ۸ درصد بوده است. این اختلاف در انتخاب نمایه می‌تواند در تفاوت دیدگاه بین نویسنده و نمایه‌ساز بوده باشد. از طرفی اختلاف این دو نشان‌دهنده جامع و کامل بودن مجموعه نمایه‌های انتخاب شده باشد.

جدول ٤-٢ میزان کل نمایه‌هایی که توسط نویسنده و نمایه‌ساز مورداستفاده قرارگرفته‌اند

	کل کلمات	نمایه‌ساز	نویسنده	میزان اشتراک
کلمات یکتا	۳۸۳۳	۲۵۶۰	۱۵۷۳	۳۰۰ کلمه = ٪۸
کل کلمات	۵۵۸۳	۳۸۶۴	۱۷۱۹	–
میانگین کلمات یکتا در هر رساله	۷,۲۷	۴,۸۶	۲,۹۸	–
میانگین کل کلمات در هر رساله	۱۰,۵۹	۷,۳۳	۳,۲۶	–

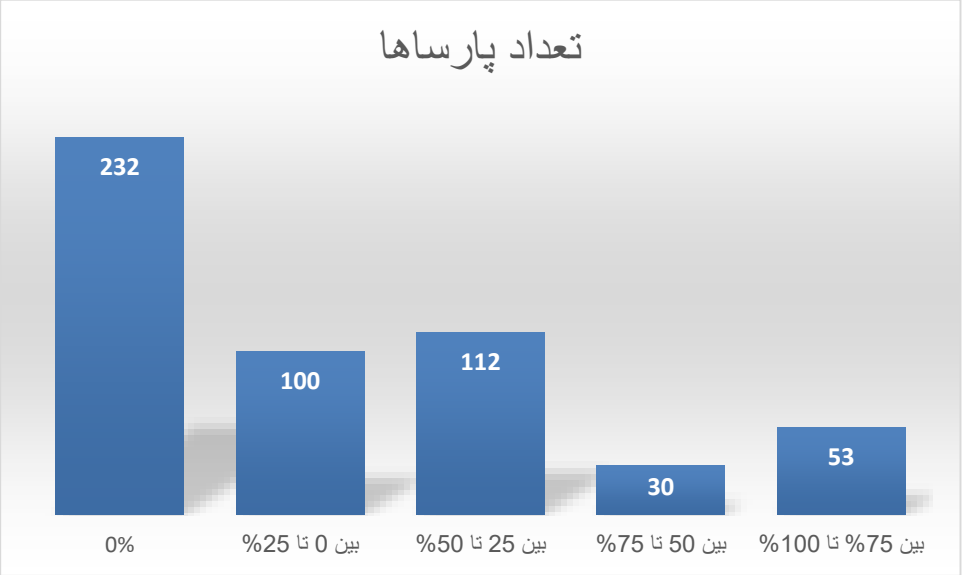
معیار دیگری که برای بازیابی بهتر اسناد در برخی انتشارات مانند الزویر ، پروکست و امرالد برای انتخاب کلیدواژگان پیشنهادشده است، انتخاب کلماتی خارج از عنوان برای نمایه است. برای سنجش این پارامتر کلیدواژگان که توسط نویسنده و نمایه‌ساز انتخاب‌شده‌اند را با کلمات عنوان

پارسا مقایسه شده‌اند. جدول ۴-۳ نشان می‌دهد که حدود ۷۶ درصد از کلیدواژه‌ها عینا در عنوان آمده‌اند (بدون ریشه‌یابی و نرمال سازی).

موضوع دیگری که بررسی می‌شود اینست که به چه میزان کلیدواژگان انتخابی دانشجو در عنوان پارسا قرار دارند؟ جدول ۴-۳ نشان می‌دهد که از بین ۵۲۷ پارسا موردبررسی حدوداً کلیدواژه‌های ۵۳ پارسا (۱۰٪ پارساها) با عنوان انطباق کامل داشته‌اند و در ۲۳۲ پارسا، کلیدواژه‌ها آن‌ها از عنوان استخراج‌نشده است (۴۴٪ از پارساها کلا از عنوان کلیدواژه‌ای انتخاب نکرده‌اند). از طرفی دیگر در ۱۳۸ پارسا هیچ توصیفگری از عنوان استخراج‌نشده است و هیچ یک از پارسا توصیفگرهای آن کلاً از عنوان استخراج‌نشده است، جدول ۴-۴. از طرفی دیگر در برخی انتشارات مانند الزویر ، پروکست و امرالد برای انتخاب کلیدواژگان پیشنهادشده است، انتخاب کلماتی خارج از عنوان برای نمایه است. بنابراین می‌توان نتیجه گرفت از آنجاکه عنوان‌ها شامل کلیدواژه‌هایی هستند که در نمایه‌ها به آن‌ها اشاره نشده است می‌بایست برای تحلیل‌روند علاوه بر نمایه‌ها از کلیدواژه‌های عنوان نیز استفاده کرد.

جدول ٤-۳ میزان اشتراک کلیدواژه‌های نویسنده با عنوان

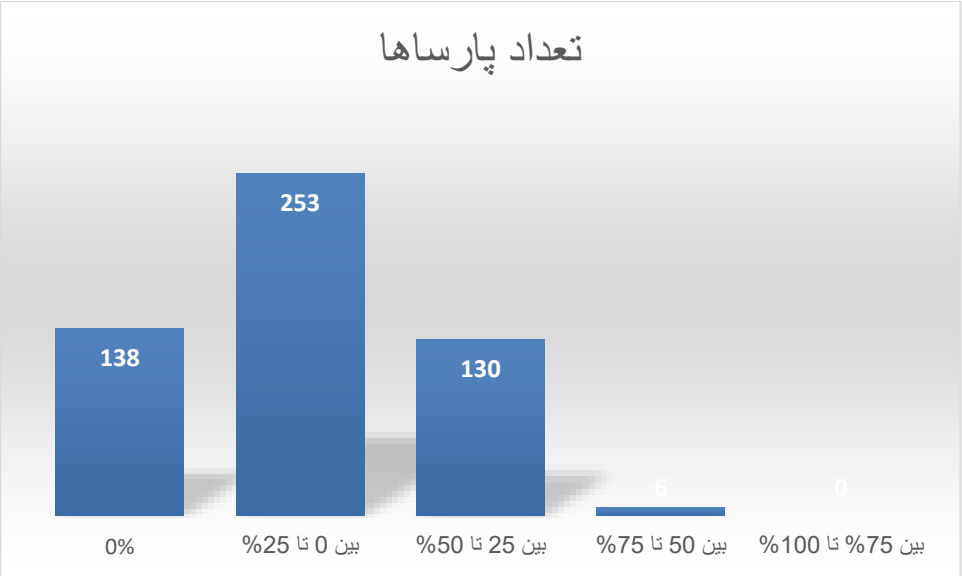
میزان اشتراک با عنوان	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارساها	۲۳۲	۱۰۰	۱۱۲	۳۰	۵۳



شکل ٤-١میزان اشتراک کلیدواژه‌های نویسنده با عنوان

جدول ٤-٤ میزان اشتراک توصیفگرها با عنوان

میزان اشتراک با عنوان	٪ ۰	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارساها	۱۳۸	۲۵۳	۱۳۰	۶	۰

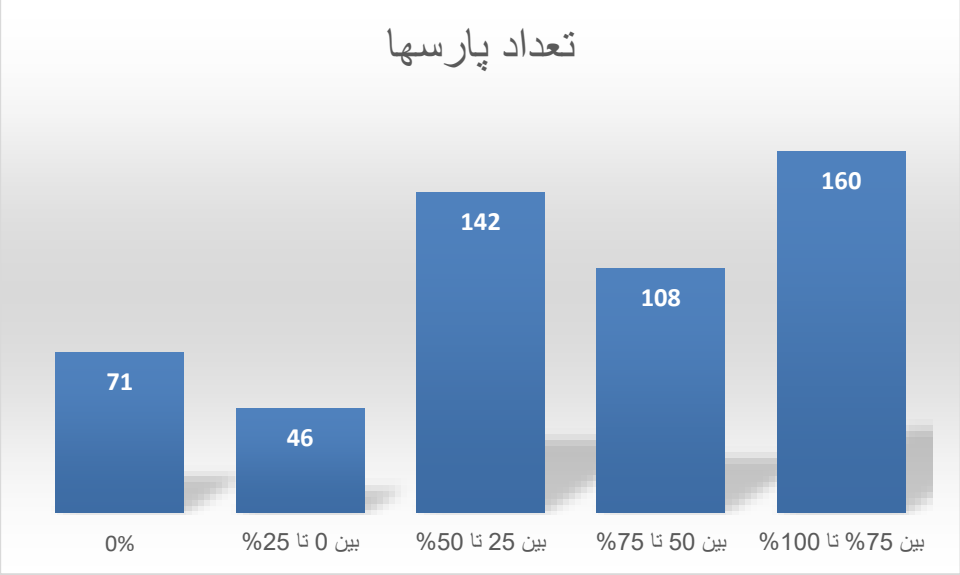


شکل ٤-٢میزان اشتراک توصیفگرها با عنوان

برای تحلیل دقیق‌تر میزان انطباق کلیدواژه‌ها و توصیفگرها با عنوان پارسا و چکیده پارسا موردبررسی قرارگرفته‌اند، که مشخص می‌کند نویسنده و نمایه‌ساز چه میزان از کلمات داخل چکیده و عنوان استفاده کرده‌اند. بررسی انجام‌شده نشان می‌دهد که در ۱۶۰ نمونه پارسا تمامی کلیدواژه‌ها از چکیده و عنوان استفاده‌شده است (حدود ۳۰٪ پارساها). از طرفی جدول ۷ نشان می‌دهد که در ۵۰٪ از پارساها بیشتر کلیدواژه‌ها از داخل چکیده و عنوان استخراج‌شده‌اند. در جدول ۸ نشان داده‌شده است که میزان انطباق توصیفگرها هم با چکیده و هم عنوان ۱٪ است و ۶٪ از پارساها کلاً توصیفگرهایشان در چکیده و در عنوان نبوده است.

جدول ٤-٥ میزان اشتراک کلیدواژه‌های نویسنده با عنوان و چکیده

میزان اشتراک	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارسها	۷۱	۴۶	۱۴۲	۱۰۸	۱۶۰

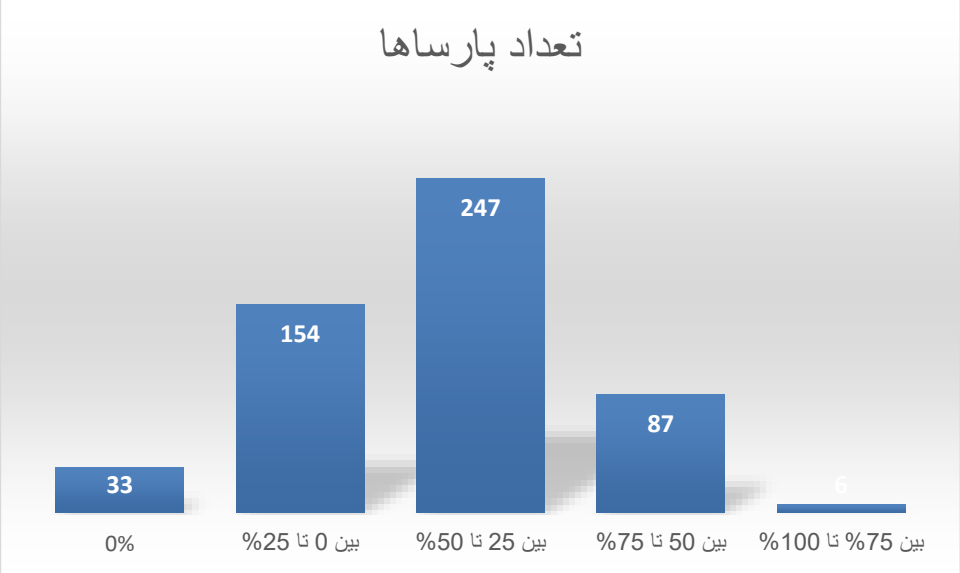


شکل ٤-٣ میزان اشتراک کلیدواژه‌های نویسنده با عنوان و چکیده

جدول ٤-٦ میزان اشتراک توصیفگرها با عنوان و چکیده

میزان اشتراک	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪

تعداد پارساها	۳۳	۱۵۴	۲۴۷	۸۷	۶



شکل ٤-٤ میزان اشتراک توصیفگرها با عنوان و چکیده

تحلیل دیگری که در این پژوهش انجام شده است بررسی و محاسبه توزیع کلمات‌کلیدی انتخاب شده توسط دانشجو و مکان رخداد آن کلمات در صورت وجود در متن چکیده رساله است. طبق استاندارد انسی/نیزو[1] چکیده شامل ۵ بخش است که به ترتیب عبارتند از: هدف چکیده؛ روش‌شناسی؛ نتایج؛ جمع‌بندی و اطلاعات اضافی. بنابراین با بررسی توزیع کلیدواژه‌ها در چکیده می‌توان حدودا به میزان اهمیت هر بخش پی‌برد و احتمال وجود یک کلیدواژه در بخش‌های مختلف چکیده را بررسی نمود. در این تحقیق چکیده را به پنج قسمت مساوی تقسیم کرده‌ایم و تعداد رخدادهای کلیدواژه‌ها را در هر بخش که توسط دانشجو و نمایه‌ساز انتخاب شده‌اند را محاسبه نمودیم. در نهایت به طور کلی احتمال وقوع رخداد کلیدواژه‌هایی که توسط دانشجو و نمایه‌ساز انتخاب شده‌اند را در هر قسمت ازچکیده به صورت جداگانه محاسبه و آن‌ها را در جدول ۴-۷ ترسیم نمودیم.

[1] استاندارد ANSI/NISO Z ۱۴.۳۹ ، استانداری است که برای چکیده نویسی ارائه شده است. در این استاندارد برای انواع متون ساختار چکیده ارائه گردیده است تا نویسندگان در نوشته‌های خود از آن استفاده کنند.

جدول ٤-٧ میزان اشتراک توصیفگرها و کلیدواژه‌های نویسنده با عنوان و چکیده

	درصد کلیدواژه‌ها	درصد توصیفگرها
در ۲۰٪ اول چکیده قرار دارند	٪ ۴۰	٪ ۴۵
در ۲۰٪ دوم چکیده قرار دارند	٪ ۲۴	٪ ۱۹
در ۲۰٪ سوم چکیده قرار دارند	٪ ۱۸	٪ ۱۸
در ۲۰٪ چهارم چکیده قرار دارند	٪ ۱۱	٪ ۱۰
در ۲۰٪ پنجم چکیده قرار دارند	٪ ۷	٪ ۷

در تحلیل دیگر بیشتر علاوه بر رشته‌های مهندسی ذکر شده از ۹۹۰ پارسا رشته‌های غیر مهندسی در بازه زمانی ذکر شده استفاده کردیم تا ۵ معیار زیر را مورد بررسی قرار دهیم.

- طول چکیده عنوان برحسب تعداد کلمه؛
- طول عنوان برحسب تعداد کلمه؛
- میانگین تعداد کلمات‌کلیدی که دانشجو برای متن انتخاب کرده‌اند؛
- میانگین تعداد کلمات‌کلیدی که نمایه‌ساز برای متن انتخاب کرده‌اند؛
- میزان اشتراک کلمات انتخاب شده توسط دانشجو و نمایه‌ساز.

این معیارها به تفکیک رشته‌های مهندسی و غیرمهندسی در جدول ۴-۸ آورده شده‌اند.

جدول ٤-٨ برخی شاخص‌های استخراج شده از پارسا به تفکیک مهندسی و غیرمهندسی

رشته	میانگین طول چکیده (کلمه)	میانگین طول عنوان (کلمه)	میانگین تعداد کلیدواژه‌های دانشجو	میانگین تعداد کلیدواژه‌های نمایه‌ساز	میانگین اشتراک کلیدواژه‌های نمایه‌ساز و دانشجو
مهندسی	۳۰۸	۱۳/۵	۳/۳	۸	۲/۶
غیرمهندسی	۳۰۱	۱۵	۳/۲	۷/۵	۲/۴

از طرف دیگر رابطه همبستگی بین تعداد کلیدواژه‌های انتخابی توسط دانشجو و طول چکیده و عنوان پارسا موردبررسی قرار گرفت. این بررسی نشان داد که هیچ رابطه‌ی معناداری بین طول

چکیده و طول عنوان با تعداد کلیدواژه‌ها انتخاب‌شده توسط دانشجو وجود ندارد و همبستگی آن‌ها برابر صفر است. برای بررسی میزان همبستگی ابتدا بردارهای کلیدواژه و طول عنوان را با استفاده از برنامه متلب ضریب همبستگی پیرسون آن‌ها را محاسبه کرده که بسیار نزدیک به ۰ شد. در محاسبه همبستگی درصورتی‌که مقدار دو بردار برابر ۱ شود یعنی آن دو بردار با یکدیگر رابطه مستقیم و درصورتی‌که برابر ۱- شوند بدان معنیست که رابطه معکوس دارند. درصورتی‌که همبستگی دو بردار برابر ۰ شود بدان معنیست که هیچ رابطه‌ای بین آن دو بردار وجود ندارد. همبستگی بین بردار تعداد کلیدواژه‌ها و طول عنوان نیز برابر صفر شد.

۴-۲- انتخاب کلیدواژه

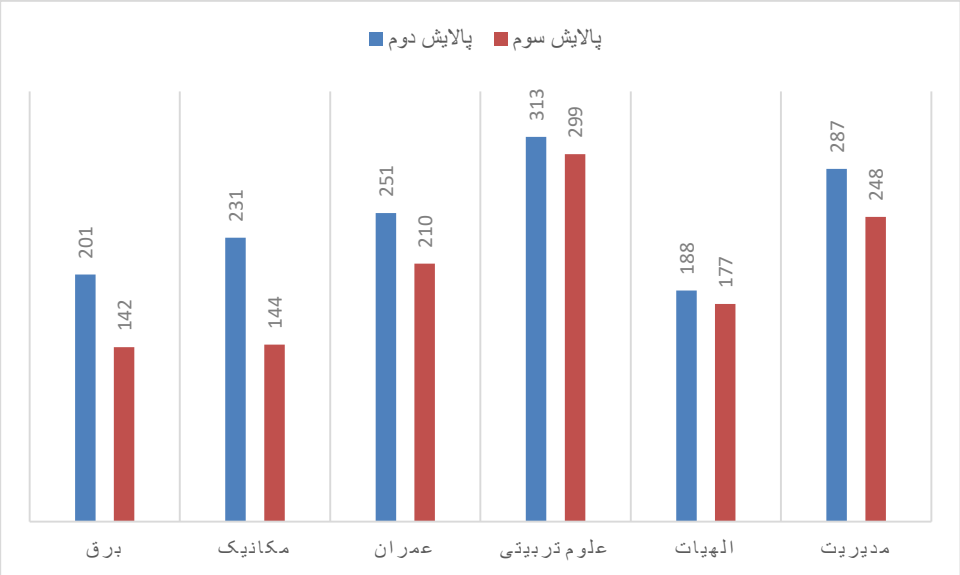
همانطور که در فصل سوم در شکل ۳-۲مراحل آماده‌سازی مجموعه‌داده‌ها نشان‌داده شده است، برای انتخاب کلیدواژه از مراحل مختلفی استفاده شده است. از جمله این گام‌ها حذف تمامی افعال، قیدها و کلمات ربط است که این پالایش توسط ابزار پردازش زبان طبیعی دانشگاه فردوسی مشهد انجام شده است. در یکی دیگر از گام‌ها حذف کلمات‌ایست از بین کلمات استخراج شده کاندید است. در این گام ابتدا از کلمات‌ایست استخراج شده در (Davarpanah, Sanji, and Aramideh ۲۰۰۹) استفاده شده است که شامل ۳۳۶ کلمه است.

بعد از پالایش کلمات کاندید و حذف کلمات‌ایست اولیه، نوبت به پالایش و حذف کلمات کم تکرار و کلمات پر تکرار آن حوزه (که کلمات‌ایست جدید محسوب می‌شوند) خواهد است. کلمات کم تکرار در این رساله کلماتی حساب می‌شوند که کمتر از ۵ ٪ پارسا شرکت داشته باشند و کلمات پرتکرار کلماتی هستند که در بیش از نصف پارسا موجود باشند که این کلمات جز کلمات‌ایست ثانویه قرار می‌گیرند. کلماتی که در کمتر از ۵ درصد پارسا قرار داشته باشند نشان دهنده حوزه‌ای نخواهند بود و مناسب برای تحلیل‌روند نیستند. کلمات پرتکرار ذکر شده به صورت کلمات کلی آن حوزه در نظر گرفته شده‌اند که باعث تمایز بین حوزه‌ها نخواهند شد، لذا این کلمات به کلمات‌ایست ثانویه اضافه شده‌اند. در پالایش سوم استفاده از روش انتخاب کلیدواژه‌های بیان شده در بخش ۳-۳ است که برای پیدا کردن مکانی که شیب کلیدواژه‌ها به صورت یکنواخت و

ثابت قرار می‌گیرد، بعد با قرار دادن پارامتر m در رابطه ۳-۳ برابر ۹ و پارامتر T برابر ۵ تعدادی کلیدواژه نیز حذف خواهند شد.

جدول ٤-٩ پالایش کلیدواژه‌ها برای تک رشته‌ای

کلمات (یکتا)	برق	مکانیک	عمران	علوم‌تربیتی	الهیات	مدیریت
کاندیدهای اولیه	۲۳۵۷۵	۲۰۱۸۸	۲۲۵۶۶	۱۱۳۰۸	۱۳۴۶۸	۱۵۵۴۴
کلمات‌ایست ثانویه	۲۳۳۷۱	۱۹۹۵۷	۲۲۳۱۵	۱۰۹۹۵	۱۳۲۸۰	۱۵۲۵۷
پالایش دوم	۲۰۱	۲۳۱	۲۵۱	۳۱۳	۱۸۸	۲۸۷
پالایش سوم	۱۴۲	۱۴۴	۲۱۰	۲۹۹	۱۷۷	۲۴۸
درصد اختلاف	٪ ۳۰	٪ ۳۷	٪۱۶	٪۴	٪ ۶	٪ ۱۴

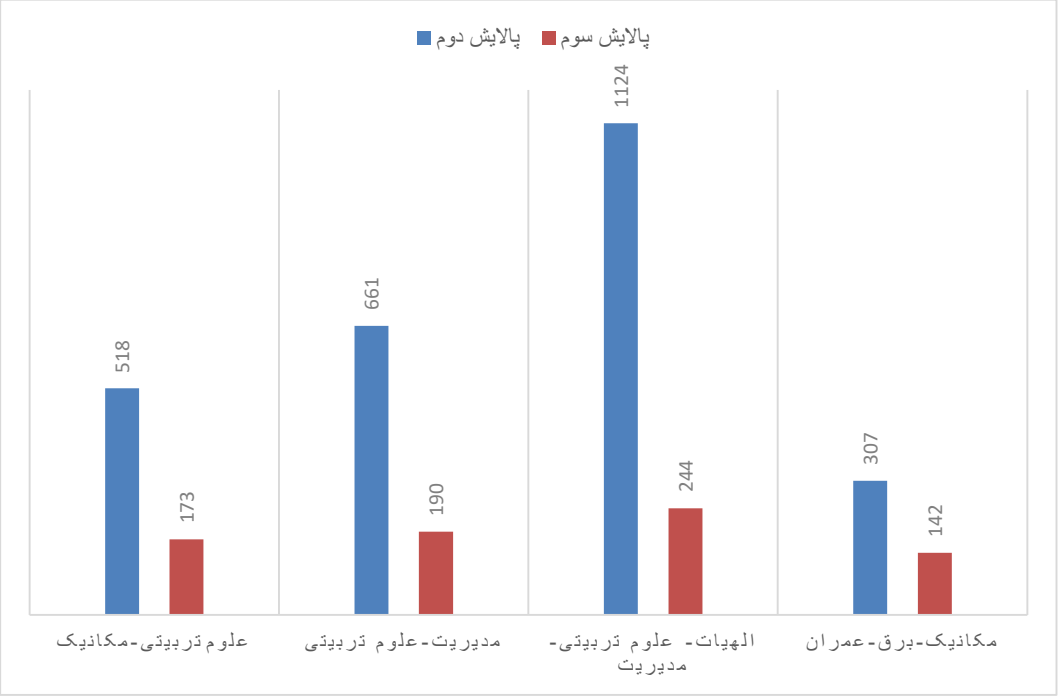


شکل ٤-٥ تعداد کلیدواژه‌های انتخاب شده با استفاده از روش پیشنهادی در مراحل پالایش برای داده‌های چند رشته‌ای

در زمانی که چند رشته را ترکیب می‌کنیم با استفاده رابطه ۳-۳ و ۳-۴ میزان فشرده‌شدگی بیشتری را به همراه خواهد داشت که جدول ۴-۱۰ پالایش کلیدواژه‌ها برای ترکیب رشته‌ها را نشان می‌دهد. همانطور که در شکل ۴-۶ ملاحظه می‌کنید در پالایش سوم تعداد کلیدواژه‌ها به صورت چشمگیری کاهش یافته است. این کاهش در تعداد کلیدواژه‌ها (بین ۵۰ تا ۸۰ درصدی) تاثیر زیادی در کاهش درستی دسته‌بندی با استفاده از کلیدواژه*های* پالایش‌شده نگذاشته است (حداکثر دو درصد درستی دسته‌بندی کاهش پیدا کرده است).

جدول ٤-۱۰ پالایش کلیدواژه‌ها برای ترکیب رشته‌ها و دسته‌بندی اسناد بر اساس کلیدواژه‌های هر مرحله از پالایش

کلمات (یکتا)	علوم‌تربیتی-مکانیک	مدیریت-علوم تربیتی	الهیات-علوم تربیتی-مدیریت	مکانیک-برق-عمران
پالایش دوم	۵۱۸	۶۶۱	۱۱۲۴	۳۰۷
صحت دسته‌بند پارامتر 1F	۹۹/۷	۹۳/۴	۹۴/۴	۸۷/۹
پالایش سوم	۱۷۳	۱۹۰	۲۴۴	۱۴۲
صحت دسته‌بند پارامتر 2F	۹۹/۶	۹۱/۲	۹۱/۷	۸۰/۵
درصد اختلاف در تعداد کلیدواه‌ها	٪۶۶	٪ ۷۱	٪ ۷۸	٪ ۵۴
درصد اختلاف در دسته‌بتدی	۰/۱	۲/۲	۲/۷	۷/۴



شکل ٤-٦ تعداد کلیدواژه‌های انتخاب شده با استفاده از روش پیشنهادی در مراحل پالایش برای داده‌های چند رشته‌ای

۴-۳- دسته‌بندی اسناد و ارزیابی آن‌ها

ازآنجاکه در فاز نخست روش پیشنهادی برای تحلیل‌روند می‌بایست اسناد علمی در حوزه‌های مختلف را از هم تفکیک کنیم، بنابراین می‌بایست از روش‌های دسته‌بندی اسناد برای تفکیک آن‌ها استفاده شود. ازاین‌رو در این بخش به بررسی و ارزیابی روش‌های موجود برای تفکیک اسناد خواهیم پرداخت. آمار هرکدام از این رشته‌ها به تفکیک در جدول ۴-۱۱ آمده است:

جدول ٤-١١ کل رشته‌های موردبررسی در آزمایش‌ها به تفکیک هر رشته

نام رشته	الهیات	علوم تربیتی	مدیریت	مهندسی مکانیک	مهندسی عمران	مهندسی برق	مجموع
تعداد	۹۵۵۸	۶۴۲۸	۴۶۳۳	۴۴۳۸	۵۵۰۸	۵۱۷۷	۳۵۷۴۴

برای انجام آزمایش‌های مختلف، ترکیب‌های مختلف از این چند مجموعه فراداده‌ها ساخته‌شده و بر روی هرکدام آزمایش‌های مربوطه انجام‌شده و مورد تحلیل و بررسی قرار می‌گیرند.

مجموعه‌داده‌های ساخته‌شده در هر بخش به تفکیک بیان‌شده‌اند. روش‌های دسته‌بندی‌های گوناگونی وجود دارد. این روش‌های خود به روش‌های عملکردی متفاوتی دارند و بر اساس آن به دسته‌های متفاوتی تقسیم شده‌اند. برای جامعیت و تحلیل بیشتر از معروفترین دسته‌بندی‌های هر گروه به عنوان نماینده آن گروه استفاده شده است. دسته‌بندی‌هایی که برای ارزیابی مورداستفاده قرارگرفته‌اند به‌قرار زیر است:

- دسته‌بند بیزین[۱] از انواع دسته‌بندهای بیزین؛
- دسته‌بند جنگل تصادفی[۲] از انواع دسته‌بندهای درختی؛
- دسته‌بندهای ترکیبی بگینگ[۳] از انواع دسته‌بندهای فرا اکتشافی؛
- مدل‌های رگرسیون لجستیک[۴] از انواع روش‌های رگرسیون به‌عنوان دسته‌بند.

برای ارزیابی نتایج روش‌های دسته‌بندی فوق از سه معیار ارزیابی دقت، بازیابی و صحت (معیار F-۱) استفاده خواهد شد (برای آشنایی این معیارها به فصل دوم مراجعه شود). بنابراین پارامترهای مورداندازه‌گیری برای ارزیابی دسته‌بندی به‌قرار زیر است:

- دقت[۵]
- بازیابی[۶]
- صحت[۷] (معیار F-۱)

در نتایج آموزش دسته‌بندی‌ها برای پیدا نشدن مشکل «بیش برازش[۸]» از روش «اعتبارسنجی متقاطع»[۹] استفاده می‌کنیم. یکی از روش‌های «اعتبارسنجی متقاطع» روش ۱۰fold- است. در ایت

---

[۱] Bayesian Network (BN)

[۲] Random Forest (RF)

[۳] Bagging (B)

[۴] Logistic Regression Models (LR)

[۵] Precision

[۶] Recall

[۷] Accuracy F-۱ Measure

[۸] Over Fitting

[۹] Cross Validation

روش داده‌های نمونه را به ۱۰ قسمت تقسیم کرده و در هر بار اجرا یکی از این بخش‌ها را به‌عنوان مجموعه‌داده آزمایش و مابقی برای آموزش مورداستفاده قرار می‌گیرد.

در آزمایش اول دسته‌بندی بر روی رشته‌های علوم انسانی (سه رشته تحصیلی استخراج‌شده از پایگاه داده گنج) انجام‌شده است. درمجموع ۲۰ هزار پایان‌نامه برای دسته‌بندی مورداستفاده قرارگرفته است. آمار هرکدام از این رشته‌ها به تفکیک در جدول ۴-۱۲ آمده است:

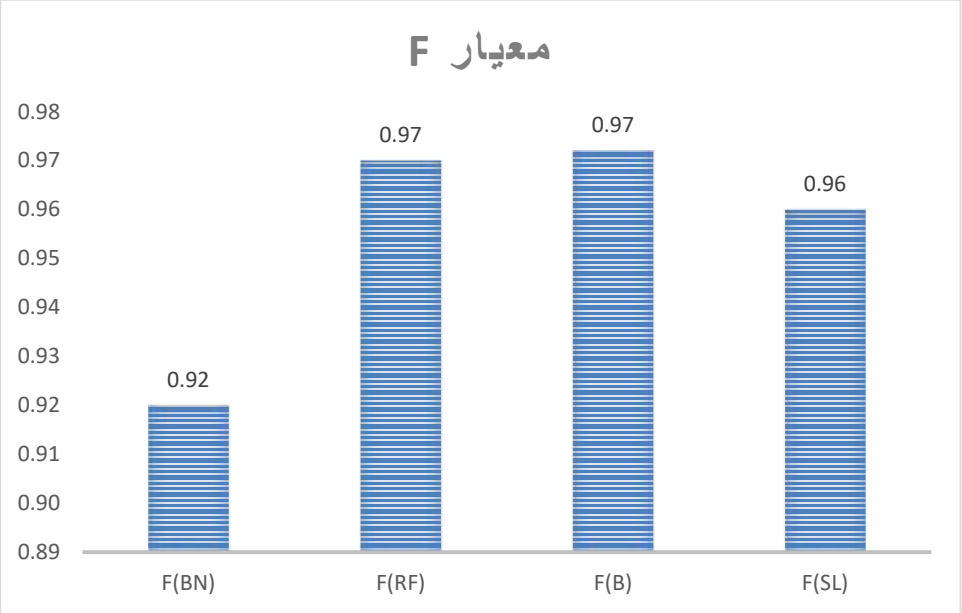
جدول ٤-۱۲ تعداد اسناد مربوط به سه رشته علوم انسانی در آزمایش اول

نام رشته	الهیات	علوم تربیتی	مدیریت	مجموع
تعداد	۹۵۵۸	۶۴۲۸	۴۶۳۳	۲۰۶۱۹

با استفاده از روش بیان‌شده در فصل روش پژوهش برای سه رشته فوق، تعداد ۲۴۴ کلیدواژه به‌عنوان ویژگی استخراج‌شده است که به‌طور متوسط برای هر ۸۵ سند یک کلیدواژه انتخاب‌شده است.

جدول ٤-۱۳ دسته‌بندی پارساهای سه رشته علوم انسانی استخراج‌شده از سامانه گنج با استفاده از ارزیابی تقاطعی ۱۰fold-

رشته	F(BN)	F(RF)	F(B)	F(SL)	R(BN)	R(RF)	R(B)	R(SL)	P(BN)	P(RF)	P(B)	P(SL)
اول	٪۹۶	٪۹۹	٪۹۴,۹	۹۴,۵	٪۹۸	٪۱۰۰	٪۹۳,۴	٪۹۴,۱	٪۹۴	٪۹۸	٪۹۶,۴	٪۹۵
دوم	٪۸۸	٪۹۷	٪۹۶,۵	۹۵,۴	٪۸۴	٪۹۸	٪۹۶,۹	٪۹۴,۸	٪۹۲	٪۹۳	٪۹۶,۲	٪۹۶
سوم	٪۸۸	٪۹۵	٪۹۸,۹	۹۸,۸	٪۹۰	٪۹۱	٪۹۹,۴	٪۹۹,۵	٪۸۶	٪۹۹	٪۹۸,۴	٪۹۸,۲
مجموع	٪۹۲	٪۹۷	٪۹۷,۲	۹۶,۸	٪۹۲	۹۷,۲	٪۹۷,۳	٪۹۶,۸	٪۹۲	٪۹۷	٪۹۷,۲	٪۹۶,۸



شکل ٤-٧ مقایسه معیار F برای دسته‌بندی سه رشته علوم انسانی با استفاده از چهار روش دسته‌بندی

جدول ۴-۱۳ نشان می‌دهد که با استفاده از کلیدواژه‌های استخراج‌شده، الگوریتم‌های دسته‌بندی توانسته‌اند با دقت بالایی سه رشته از یک حوزه (علوم انسانی) را از هم تفکیک کنند. در این جدول همچنین نشان داده‌شده است که با استفاده از الگوریتم جنگل تصادفی می‌توانیم به بهترین نتیجه در بین مابقی روش‌ها برای دسته‌بندی برسیم.

در آزمایش دوم برای ارزیابی و بررسی بیشتر روش‌های دسته‌بندی را بر روی سه رشته مهندسی (مکانیک، عمران و برق) موجود در پایگاه داده انجام خواهیم داد. این پارساها با استفاده از الگوریتم‌ها و پارامتری بیان‌شده در فوق مورد ارزیابی قرارگرفته‌اند. همان‌طور که در جدول ۴-۱۴ نشان داده‌شده است درمجموع ۱۵۱۲۵ پارسا در این آزمایش قرار دارند که درمجموع با استفاده از روش بیان‌شده در فصل سوم ۲۶۳ کلیدواژه به‌عنوان ویژگی برای تفکیک این سه دسته از هم انتخاب‌شده است. که به طور متوسط برای هر ۵۸ پارسا یک کلیدواژه انتخاب‌شده است. آمار تعداد پارساهای این آزمایش به تفکیک در جدول ۴-۱۴ نشان داده‌شده است.

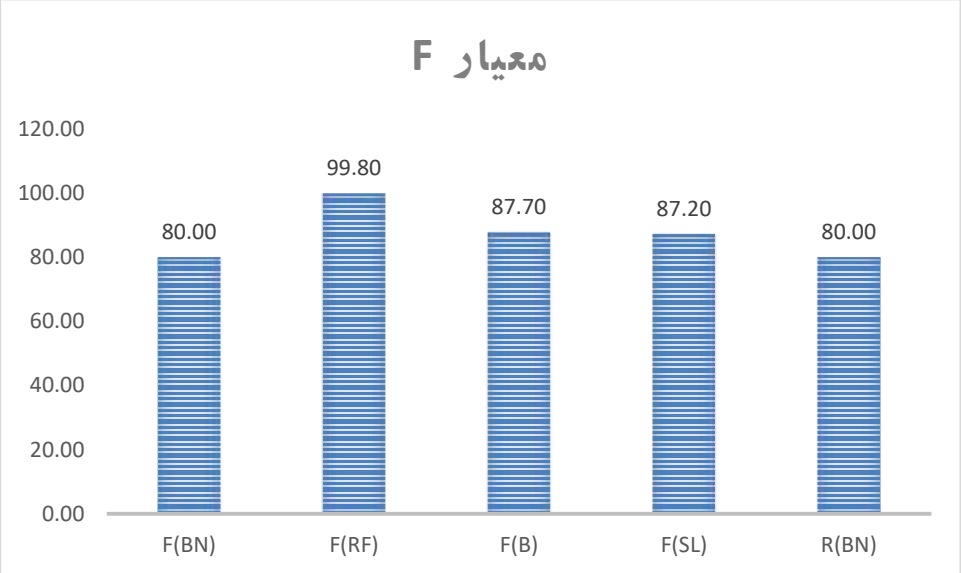
جدول ٤-١٤ سه رشته مهندسی موردبررسی به تفکیک هر رشته

نام حوزه	مهندسی مکانیک	مهندسی عمران	مهندسی برق	مجموع
تعداد اسناد	۴۴۳۸	۵۵۰۸	۵۱۷۷	۱۵۱۲۵

جدول ۴-۱۵ نشان می‌دهد که با استفاده از کلیدواژه‌های استخراج‌شده برای سه رشته از یک حوزه متفاوت دیگر (مهندسی)، الگوریتم‌های دسته‌بندی توانسته‌اند با دقت بالایی آن‌ها را از هم تفکیک کنند. الگوریتم جنگل تصادفی نیز از بقیه الگوریتم‌ها دقت بالاتری به ما می‌دهد.

جدول ٤-١٥دسته‌بندی پارساهای سه رشته مهندسی استخراج‌شده از سامانه گنج با استفاده از ارزیابی تقاطعی ۱۰-fold

رشته	F(BN)	F(RF)	F(B)	F(SL)	R(BN)	R(RF)	R(B)	R(SL)	P(BN)	P(RF)	P(B)	P(SL)
اول	٪۸۳	۹۹,۸	۸۳,۹	۸۳,۳	٪۹۰	۹۹,۹	۸۰,۴	۸۰,۹	٪۷۷	۹۹,۷	۸۷,۷	۸۵,۸
دوم	٪۷۴	۹۹,۸	۸۹,۵	۸۹,۲	٪۷۱	۹۹,۸	۸۸,۹	۸۶,۷	٪۷۸	۹۹,۹	۹۰,۲	۹۱,۹
سوم	٪۸۳	۹۹,۷	۸۹	۸۸,۴	٪۷۹	۹۹,۷	۹۳	۹۳,۲	٪۸۷	۹۹,۷	۸۵,۴	۸۴,۰
مجموع	٪۸۰	۹۹,۸	۸۷,۷	۸۷,۲	٪۸۰	۹۹,۸	۸۷,۸	۸۷,۲	٪۸۰	۹۹,۸	۸۷,۹	۸۷,۴



شکل ٤-٨ مقایسه معیار F برای دسته‌بندی سه رشته مهندسی با استفاده از چهار روش دسته‌بندی

در آزمایش سوم در این بخش دو حوزه مهندسی و علوم انسانی را باهم ترکیب کرده و مجموعه‌داده جدید متشکل از چهار رشته مکانیک، عمران، تربیت‌بدنی، و مدیریت، مجموعه‌داده‌ای درست کردیم. برای هر یک از رشته‌ها تعداد اسناد موجود به شرح جدول ۴-۱۶ است. درمجموع

۲۱ هزار سند با ۲۹۰ ویژگی تشکیل‌شده است. که به‌صورت متوسط برای هر ۷۲ سند یک کلیدواژه انتخاب‌شده است.

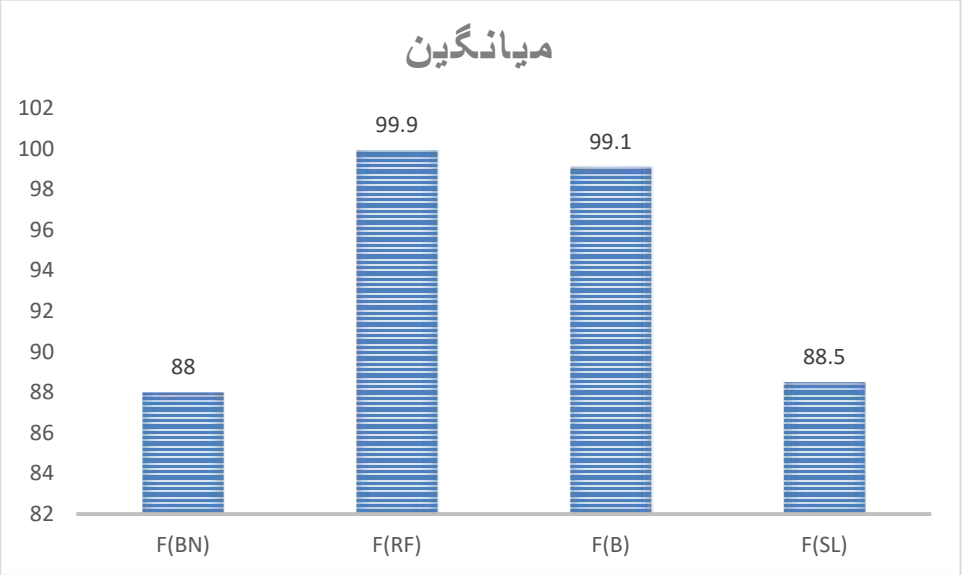
جدول ٤-١٦تعداد رشته‌های موردبررسی به تفکیک هر رشته

نام حوزه	مهندسی مکانیک	مهندسی عمران	مدیریت	تربیت‌بدنی	مجموع
تعداد اسناد	۴۴۳۸	۵۵۰۸	۴۶۳۳	۵۶۲۸	۲۱۰۰۷

جدول ۴-۱۷ نشان می‌دهد که با استفاده از کلیدواژه‌های استخراج‌شده برای سه رشته از یک حوزه متفاوت دیگر (مهندسی)، الگوریتم‌های دسته‌بندی توانسته‌اند با دقت بالایی آن‌ها را از هم تفکیک کنند. الگوریتم جنگل تصادفی نیز از بقیه الگوریتم‌ها دقت بالاتری به ما می‌دهد.

جدول ٤-١٧دسته‌بندی پارساهای رشته‌های آزمایش سوم استخراج‌شده از سامانه گنج با استفاده از ارزیابی تقاطعی ۱۰-fold

رشته	F(BN)	F(RF)	F(B)	F(SL)	R(BN)	R(RF)	R(B)	R(SL)	P(BN)	P(RF)	P(B)	P(SL)
اول	۹۳,۱	۱۰۰	۹۵,۸	۹۳,۸	۹۲,۲	۱۰۰	۹۴,۷	۹۳,۱	۹۶,۱	۱۰۰	۹۴,۷	۹۴,۶
دوم	۸۲	۱۰۰	۹۰,۹	۸۸	۹۲,۶	۱۰۰	۹۱	۸۶,۵	۸۷,۲	۱۰۰	۹۱	۸۹,۷
سوم	۸۰	۹۹,۹	۸۸,۱	۸۵,۸	۸۲,۶	۹۹,۹	۸۸,۲	۸۴,۶	۹۱,۱	۹۹,۹	۸۸,۲	۸۷
چهارم	۸۶,۹	۹۹,۹	۸۸,۳	۸۴,۵	۹۱,۷	۹۹,۸	۸۹,۸	۸۸,۶	۸۲,۵	۹۹,۹	۸۹,۸	۸۰,۸
میانگین	۸۸	۹۹,۹	۹۹,۱	۸۸,۵	۸۹,۶	۹۹,۹	۹۱,۱	۸۸,۴	۹۰	۹۹,۹	۹۱,۱	۸۸,۶



شکل ٤-٩ مقایسه معیار F برای دسته‌بندی ترکیب رشته‌های مهندسی و علوم انسانی با استفاده از چهار روش دسته‌بندی

در انتها می‌توان نتیجه گرفت که با استفاده از الگوریتم دسته‌بندی درختی می‌توانیم با دقت بالایی اسناد مختلف را دسته‌بندی کرده و از یکدیگر تفکیک کنیم. از طرفی روش پیشنهادی برای انتخاب کلیدواژه‌ها بر اساس شیب فراوانی باعث شده که علاوه بر کاهش بالای تعداد کلیدواژه‌ها و کاهش فضای حالت روش‌های مختلف دسته‌بندی به‌خوبی بتوانند رشته‌های مختلف را از هم تفکیک کنند. برای مثال با استفاده از روش دسته‌بندی جنگل تصادفی با دقت میانگین ۹۹ درصد قابلیت تفکیک بین حوزه‌ها را دارند.

۴-۴- کشف موضوع‌های حوزه‌های علمی

همانطور که در بخش‌های ۲-۲-۵ و ۳-۴ بیان شد یکی از روش‌های کشف موضوع خوشه‌بندی ماتریس هم‌رخدادی کلیدواژه‌ها است. در این بخش با استفاده از سه روش خوشه‌بندی کا-میانه[۱]، خوشه‌بندی فازی و LDA و با استفاده از روش‌های ارائه‌شده (برای بازنمایی کلمات و تشکیل مجموعه‌داده در نمایش کلمات متون) به موضوع‌بندی حوزه‌ها خواهیم پرداخت. تا با استفاده از

[۱] K-mean

موضوع‌های تشکیل‌شده برای هر حوزه بتونیم روند آن را تعیین کنیم. در ادامه برای بررسی میزان درستی، ارزیابی و عملکرد الگوریتم ارائه‌شده با استفاده از مجموعه‌داده‌هایی که تشکیل داده‌ایم میزان خلوص خوشه‌بندی این رشته‌ها را مورد ارزیابی قرار می‌دهیم. هرچقدر میزان خلوص در خوشه‌بندی مجموعه‌داده‌ها بیشتر باشد و اسناد در خوشه‌های درستی قرار بگیرند الگوریتم پیشنهادی عملکرد بهتری نسبت به روش‌های دیگر دارد.

چندین آزمایش با استفاده مجموعه‌داده‌های متفاوتی که با استفاده از اسناد که از پایگاه داده گنج استخراج‌شده انجام‌شده است. تفکیک هر یک از مجموعه اسناد در بخش قبل ارائه‌شده است. برای ارزیابی نتایج بعد از استخراج موضوع‌ها از اسناد، دو ماتریس جدید تشکیل می‌شود: ماتریس کلمه–موضوع و ماتریس سند–موضوع. کلمه–موضوع نشان‌دهنده این است که هر موضوع از چه کلمه‌هایی تشکیل‌شده است. ماتریس سند–موضوع نشان‌دهنده این است که هر سند از چه موضوع‌هایی و به چه میزانی ساخته‌شده‌اند. حال برای بررسی عملکرد موضوع‌های استخراج‌شده با استفاده از روش‌های خوشه‌بندی و دسته‌بندی ماتریس سند–موضوع را مورد ارزیابی قرار خواهیم داد و میزان تفکیک‌پذیری اسناد با استفاده از موضوع‌های استخراج‌شده را مورد تحلیل قرار می‌دهیم. مجموعه‌داده‌هایی که برای آزمایش‌های مختلف تشکیل‌شده‌اند در جدول ۴-۱۸ آمده است. برای انجام آزمایش‌ها برای تک‌تک مجموعه‌داده‌های معرفی‌شده، ماتریس‌های هم‌رخدادی که در فصل سوم به آن اشاره شد را تشکیل داده‌شده توسط و عملکرد آن‌ها را موردبررسی قرار می‌دهیم.

جدول ٤-١٨ مجموعه‌داده‌های تشکیل‌شده برای آزمایش میزان دقت خوشه‌ها

نام مجموعه‌داده	تعداد اسناد	تعداد کلیدواژه‌ها
برق–عمران	۱۰۶۸۵	۱۹۶
مدیریت – برق	۹۱۶۵	۲۸۶
مدیریت–عمران	۱۰۱۴۱	۳۳۸
علوم‌تربیتی–مکانیک	۱۰۸۶۶	۱۷۳
علوم تربیتی–مدیریت	۱۱۰۶۱	۱۹۰
برق–مکانیک–عمران	۱۵۱۲۵	۲۶۳
علوم تربیتی–مدیریت– الهیات	۲۰۶۱۹	۲۴۴
مدیریت–علوم اجتماعی – برق – مکانیک	۲۱۰۰۷	۲۹۰

در آزمایش چهارم با استفاده از الگوریتم خوشه‌بندی کا-میانه ماتریس‌های هم‌رخدادی تشکیل‌شده از مجموعه‌داده‌های معرفی‌شده را خوشه‌بندی کرده تا موضوع‌های مختلفی تشکیل شود (در این حالت ماتریس کلیدواژه-موضوع تشکیل می‌شود). سپس با استفاده از این ماتریس، ماتریس موضوع-سند را تشکیل داده و با استفاده از الگوریتم خوشه‌بندی کا-میانه[1] پارساها را خوشه‌بندی و دسته‌بندی خواهیم کرد. هرچه میزان خلوص نتیجه خوشه‌بندی بیشتر باشد بدین معنیست که موضوع‌ها به‌خوبی تشکیل‌شده و قابلیت تفکیک یک‌رشته را از رشته دیگر دارند. از طرفی میزان درستی در دسته‌بندی نیز مورد ارزیابی قرارگرفته است. بنابراین به صورت خلاصه آزمایش‌های انجام شده به صورت زیر می‌باشند:

- تحلیل اثر روش‌های بازنمایی کلمات در خلوص خوشه‌بندی اسناد حوزه‌ها با استفاده از کا-میانه
- تحلیل اثر روش‌های بازنمایی کلمات در صحت دسته‌بندی اسناد حوزه‌ها با استفاده از کا-میانه
- تحلیل اثر روش‌های بازنمایی کلمات در خلوص خوشه‌بندی اسناد حوزه‌ها با استفاده از کا-میانه-دوبخشی
- تحلیل اثر روش‌های بازنمایی کلمات در صحت دسته‌بندی اسناد حوزه‌ها با استفاده از کا-میانه-دوبخشی
- تحلیل اثر روش‌های بازنمایی کلمات در خلوص خوشه‌بندی اسناد حوزه‌ها با استفاده از خوشه‌بندی کانوپی
- تحلیل اثر روش‌های بازنمایی کلمات در صحت دسته‌بندی اسناد حوزه‌ها با استفاده از خوشه‌بندی کانوپی

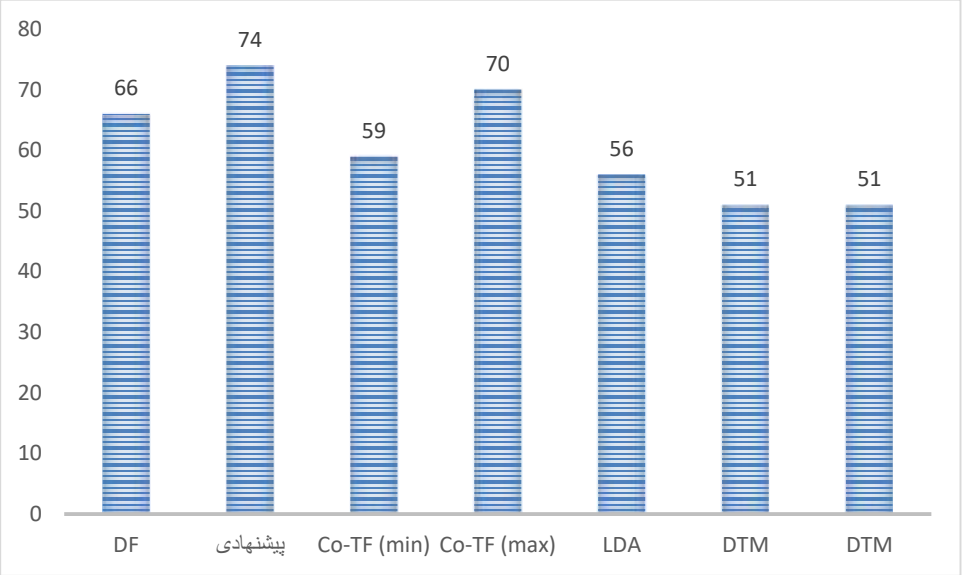
نتایج خروجی در جدول ۴-۱۹ و جدول ۴-۲۰ نشان داده‌شده است. اعداد داخل این جداول برحسب درصد نوشته‌شده‌اند. نتایج نوشته‌شده در بهترین حالت در تعداد موضوع‌ها برای هر یک از

[1] K-Means

ماتریس‌های هم‌رخدادی محاسبه‌شده و در جدول درج‌شده است. در زیر بخش بعدی به بیان جزئیات در میزان خلوص خوشه‌بندی و صحت دسته‌بندی در مجموعه‌داده‌ها خواهیم پرداخت.

جدول ٤-١٩ میزان خلوص (درست خوشه‌بندی کردن برحسب درصد) اسناد با استفاده از خوشه‌بند کا-میانه با استفاده از هر یک از شکل‌های نمایش کلمات

ردیف	نام مجموعه‌داده مورد ارزیابی	DF	پیشنهادی (Proposed)	Co-TF (min)	Co-TF (max)	LDA	DTM (tf)	DTM (tf*idf)
۱	برق-عمران	۶۰	۵۲	۵۸	۶۱	۵۱	۵۳	۵۷
۲	مدیریت – برق	۸۹	۹۶	۶۹	۸۸	۶۲	۵۲	۵۲
۳	مدیریت-عمران	۹۱	۹۴	۶۶	۹۰	۵۷	۵۸	۵۸
۴	علوم‌تربیتی-مکانیک	۵۸	۹۱	۶۴	۸۴	۵۵	۵۷	۵۴
۵	علوم تربیتی-مدیریت	۷۹	۸۲	۷۰	۸۴	۶۵	۵۴,۳	۵۵
۶	برق-مکانیک-عمران	۵۱	۵۱	۴۷	۵۰	۵۲	۴۰	۳۹
۷	علوم تربیتی-مدیریت-الهیات	۴۵	۶۴	۴۴,۵	۶۵,۵	۵۱	۵۲	۵۲
۸	مدیریت-علوم اجتماعی - برق - مکانیک	۵۲	۶۵	۵۱	۵۹	۵۲	۳۹	۳۹
۹	میانگین	۶۶	۷۴	۵۹	۷۰	۵۶	۵۱	۵۱



شکل ٤-١٠میانگین خلوص خوشه‌بندی در روش‌های مختلف بازنمایی کلمات

نتایج به‌دست‌آمده نشان می‌دهد که بدون استفاده از روش‌های ماتریس هم‌رخدادی کلمات، خوشه‌بندی اسناد از دقت پایینی برخوردار است و تقریباً به‌صورت تصادفی انجام می‌شود. اما با استفاده از ماتریس‌های هم‌رخدادی و تشکیل موضوع و ماتریس سند-موضوع نتایج بسیار متفاوت و بسیار هوشمندانه خوشه‌بندی انجام‌شده است.

جدول ۴-۱۹ نشان می‌دهد که در مجموعه‌داده‌هایی که از دو رشته مختلف در دو حوزه مختلف علوم انسانی و مهندسی تشکیل‌شده‌اند (مانند مجموعه‌داده‌های ۲، ۳ و ۴) خلوص خوشه‌بندی بسیار بالا بوده و تقریباً به تمام اسناد به‌درستی در خوشه مربوط به خود قرارگرفته‌اند. از طرفی از بین روش‌های تشکیل ماتریس هم‌رخدادی، روش پیشنهادی در ارائه ماتریس هم‌رخدادی (که در فصل قبل ارائه شد) در این رساله نتایج بهتری را نسبت به روش‌های دیگر داشته است. بعد از روش پیشنهادی استفاده از حداکثر تعداد هم‌رخدادی‌ها برای تشکیل ماتریس نتیجه بهتر را کسب کرده است.

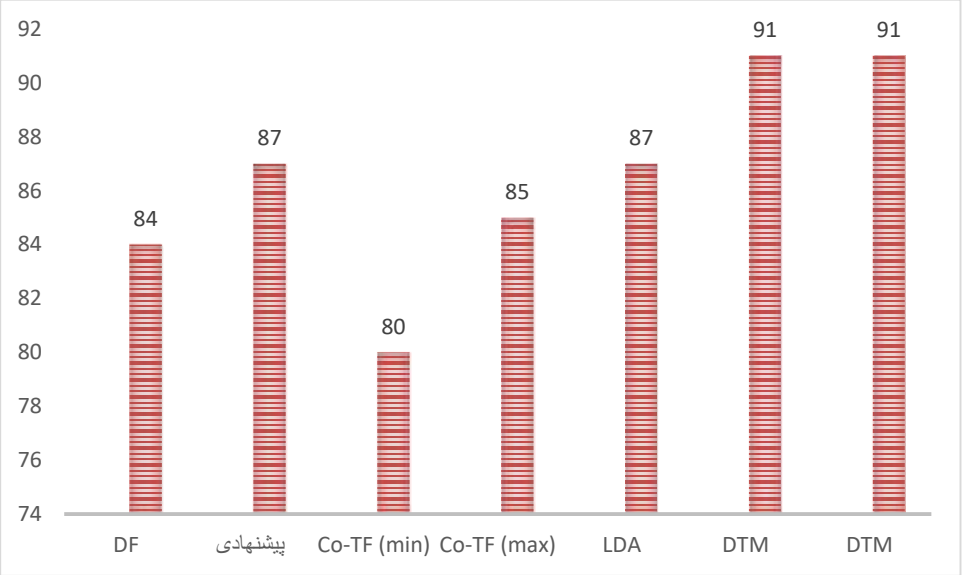
در زمانی که مجموعه‌داده‌ها از رشته‌های در یک حوزه یکسان تشکیل‌شده‌اند (مانند مجموعه‌داده‌های ۱، ۵، ۶، ۷) نتایج خلوص خوشه‌بندی کاهش پیداکرده است (اما هنوز از میزان خلوص خوشه‌بندی با استفاده از ماتریس‌های DTM با اختلاف بالاتر است). این کاهش می‌تواند به

خاطر اشتراک و شباهت در بین موضوع‌های تشکیل‌دهنده حوزه‌های علمی باشد. این خوشه‌ها موضوع‌های میان‌رشته‌ای هستند که باعث شده‌اند که یک سند در دو حوزه مختلف دسته‌بندی شود. در تفکیک ترکیب حوزه‌های مختلف باهم دیگر مانند مجموعه‌داده ۸ روش پیشنهادی در تشکیل ماتریس هم‌رخدادی بسیار بهتر از روش‌های دیگر عمل کرده است.

برای بررسی میزان تأثیر روش‌های پیشنهادی (در ارائه ماتریس هم‌رخدادی) بر میزان صحت دسته‌بندی، مجموعه‌داده‌های معرفی‌شده بعد از تشکیل ماتریس موضوع–سند با استفاده از الگوریتم دسته‌بندی بیزین مورد آزمایش قرار خواهند گرفت تا میزان درستی در دسته‌بندی انجام‌شده موردبررسی قرار گیرد. جدول ۴-۲۰ میزان صحت دسته‌بندی را برحسب درصد بیان کرده است. این نتایج نشان می‌دهند که روش پیشنهادی در تشکیل ماتریس هم‌رخدادی کلمات نتایجی بهتر از روش‌های دیگر و نزدیک به ماتریس‌های کلمه–سند دارند.

جدول ٤-۲۰ ارزیابی روش‌های پیشنهادی با استفاده از دسته‌بندی با معیار F-1 با استفاده از خوشه‌بند کا–میانه با استفاده از هر یک از شکل‌های نمایش کلمات پیشنهادی

ردیف	نام مجموعه‌داده مورد ارزیابی	DF	Proposed	Co-TF (min)	Co-TF (max)	LDA	DTM	DTM (tf*idf)
۱	برق–عمران	۷۷,۳	۸۰,۳	۷۴,۶	۷۷,۸	۸۵	۸۷	۸۷,۳
۲	مدیریت – برق	۹۷	۹۷	۹۰	۹۴,۵	۹۶	۹۸,۳	۹۸,۳
۳	مدیریت–عمران	۹۶	۹۶	۸۵,۳	۹۵,۵	۹۵	۹۷,۳	۹۷
۴	علوم‌تربیتی–مکانیک	۹۰,۷	۹۷,۱	۸۷,۵	۹۴	۹۸	۹۹,۶	۹۸,۵
۵	علوم تربیتی–مدیریت	۸۸	۸۸	۸۵	۸۸	۸۷	۹۱,۲	۹۱,۲
۶	برق–مکانیک–عمران	۶۶	۶۷	۶۱	۶۶	۷۲	۸۰,۵	۷۳,۵
۷	علوم تربیتی–مدیریت–الهیات	۸۴,۱	۸۷,۵	۸۲,۷	۸۸,۲	۸۵,۵	۹۱,۷	۹۱,۷
۸	مدیریت–علوم اجتماعی – برق – مکانیک	۷۰	۸۰	۷۲	۷۸	۸۱	۸۸	۸۸
۹	میانگین	۸۴	۸۷	۸۰	۸۵	۸۷	۹۱	۹۱



شکل ٤-١١تاثیر روش‌های بازنمایی کلمات در دسته‌بندی

در انتهای این بخش می‌توان این نتیجه را گرفت که با استفاده از روش پیشنهادی در تشکیل ماتریس‌های هم‌رخدادی علاوه بر قابلیت تشکیل موضوع با استفاده از کلیدواژه‌های اسناد می‌توان اسناد مختلف را بر اساس موضوع‌های پیداشده با دقت بسیار بالاتری نسبت به حالت بدون استفاده از این روش‌ها خوشه‌بندی کرد. مزیت دیگر استفاده از این روش در استفاده از دسته‌بندی با استفاده از موضوع‌های استخراج‌شده، دقت بالا در دسته‌بندی و نزدیک به حالت ماتریس سند-کلمه است. ازآنجاکه موضوع‌های استخراج‌شده در ماتریس سند-موضوع یک ویژگی جدید محسوب می‌شوند می‌توان روش پیشنهادشده را از نگاهی دیگر به‌عنوان روش استخراج ویژگی در نظر گرفت.

ازآنجاکه هر کلیدواژه می‌تواند در چند موضوع مختلف وجود داشته باشد و ایفای نقش کند در آزمایش‌های آتی (که در سه زیر بخش بعدی بیان‌شده است) از روش‌های خوشه‌بندی فازی (نوعی از خوشه‌بندی نرم است که در این رساله برای کشف موضوع پیشنهادشده است) و تحلیل دریکله نهان[1] برای کشف و استخراج موضوع استفاده خواهیم کرد. گونه‌ای دیگر از جایگذاری کلمات در اسناد به نام Vec۲Word نیز برای مقایسه و بررسی‌های بیشتر استفاده‌شده است. همان‌طور که در فصل دوم به آن اشاره‌شده است، این روش یکی از قدرتمندترین روش‌ها برای استخراج روابط معنایی است که مجموعه در اسناد بزرگ کارایی بسیاری دارد.

[1] Latent Dirichlet Analysis

بعد از استخراج موضوع ماتریس سند-موضوع تشکیل‌شده و میزان خلوص و صحت با جزئیات بیشتری در تعداد خوشه‌ها و حالات مختلف الگوریتم‌ها موردبررسی قرار می‌گیرد. بعد از انتخاب یک روش مناسب برای خوشه‌بندی و تشکیل ماتریس هم‌رخدادی، یک حوزه علمی انتخاب خواهد شد و آن حوزه علمی نیز به تفکیک به موضوع‌های کوچک‌تری تشکیل می‌شود. ازآنجاکه معیارهای مختلفی برای ارزیابی یک خوشه وجود دارد در ادامه (زیر بخش۴-۲-۲) به بیان روشی می‌پردازیم که با استفاده از چندین معیار ارزیابی بتوان بهترین تعداد خوشه را پیشنهاد داد.

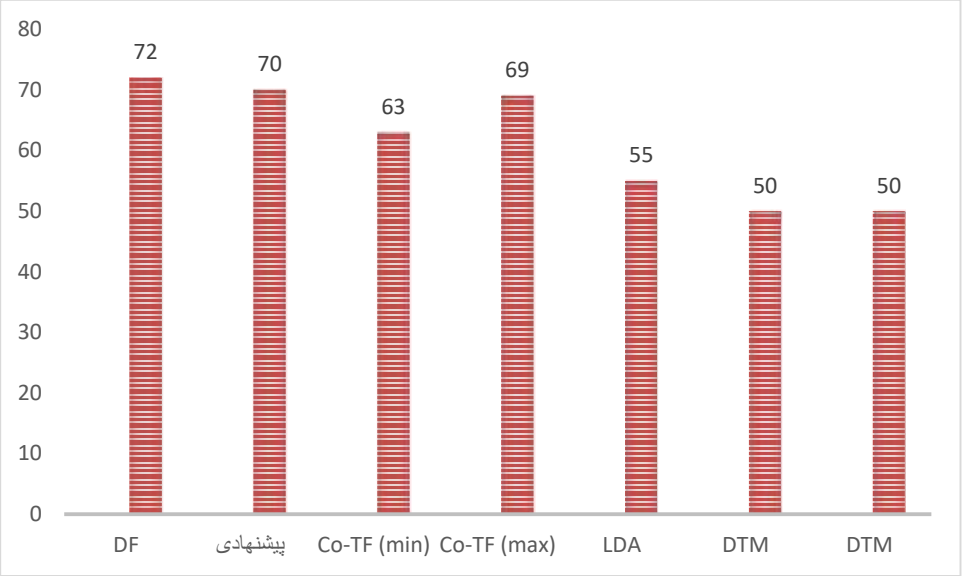
خوشه‌بندی کا-میانه-دوبخشی یکی دیگر از خوشه بندی‌های شناخته‌شده است که مطالعات در کارهای گذشته نشان داده است که عملکردی بهتر از روش‌های خوشه‌بندی کا-میانه و سلسله‌مراتبی دارد. جدول ۴-۲۱ میزان خلوص (درست خوشه‌بندی کردن برحسب درصد) اسناد با استفاده از خوشه‌بند کا-میانه-دوبخشی با استفاده از هر یک از شکل‌های نمایش کلمات را نشان می‌دهد. برخلاف نتایج کارهای گذشته، این نتایج نشان می‌دهد که عملکرد این خوشه‌بندی برای خوشه‌بندی و دسته‌بندی متون فارسی ضعیف‌تر از کا-میانه بوده است.

شکل ۴-۱۲میانگین خلوص خوشه‌بندی در روش‌های مختلف بازنمایی کلمات

جدول ۴-۲۲ ارزیابی روش‌های پیشنهادی با استفاده از دسته‌بندی با معیار F-1 با استفاده از خوشه‌بند کا-میانه-دوبخشی با استفاده از هر یک از شکل‌های نمایش کلمات پیشنهادی نشان می‌دهد.

جدول ٤-۲۱ میزان خلوص (درست خوشه‌بندی کردن برحسب درصد) اسناد با استفاده از خوشه‌بند کا-میانه-دوبخشی با استفاده از هر یک از شکل‌های نمایش کلمات

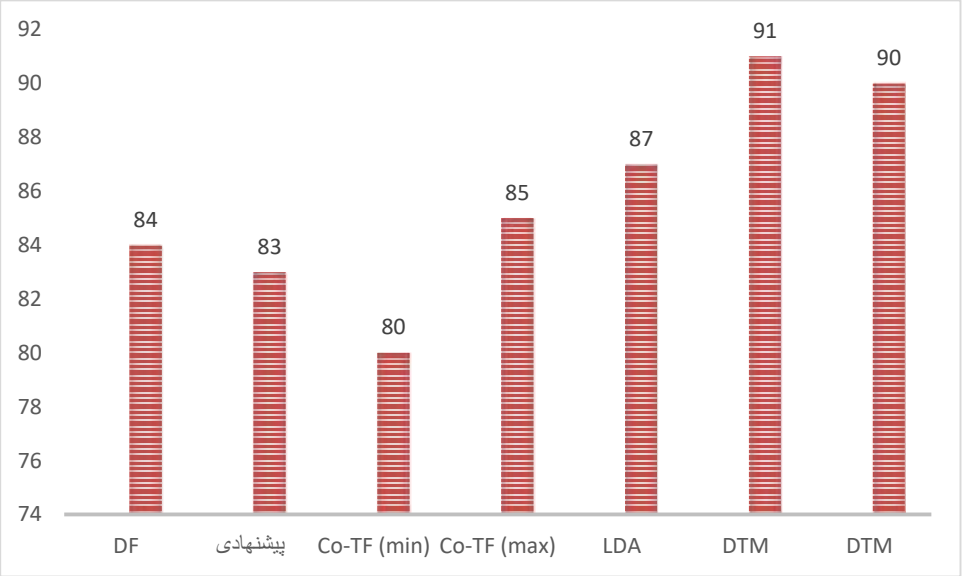
ردیف	نام مجموعه‌داده مورد ارزیابی	DF	Proposed	Co-TF (min)	Co-TF (max)	LDA	DTM	DTM (tf*idf)
۱	برق-عمران	۶۰	۵۳	۶۱	۵۸	۵۱	۵۳	۵۷
۲	مدیریت – برق	۸۸	۹۲	۷۱	۸۲	۶۲	۵۲	۵۲
۳	مدیریت-عمران	۹۳	۸۸	۷۵	۹۴	۵۷	۵۸	۵۸
۴	علوم‌تربیتی-مکانیک	۸۵	۶۴	۸۳	۷۷	۵۵	۵۷	۵۴
۵	علوم تربیتی-مدیریت	۸۳	۸۴	۶۵	۸۴	۶۵	۵۴,۳	۵۵
۶	برق-مکانیک-عمران	۵۱	۶۸	۵۳	۵۱	۵۲	۴۰	۳۹
۷	علوم تربیتی-مدیریت-الهیات	۷۴	۶۴	۶۱	۶۱	۵۱	۵۲	۵۲
۸	مدیریت-علوم اجتماعی - برق - مکانیک	۴۶	۴۹	۳۹	۵۱	۵۲	۳۹	۳۹
	میانگین	۷۲	۷۰	۶۳	۶۹	۵۵	۵۰	۵۰



شکل ٤-١٢میانگین خلوص خوشه‌بندی در روش‌های مختلف بازنمایی کلمات

جدول ٤-٢٢ ارزیابی روش‌های پیشنهادی با استفاده از دسته‌بندی با معیار F-1 با استفاده از خوشه‌بند کا-میانه-دوبخشی با استفاده از هر یک از شکل‌های نمایش کلمات پیشنهادی

ردیف	نام مجموعه‌داده مورد ارزیابی	DF	Proposed	Co-TF (min)	Co-TF (max)	LDA	DTM	DTM (tf*idf)
۱	برق-عمران	۷۹	۸۱	۷۴	۸۰	۸۵	۸۷	۸۷,۳
۲	مدیریت – برق	۹۶	۹۷	۹۰	۹۷	۹۶	۹۸,۳	۹۸,۳
۳	مدیریت-عمران	۹۵	۹۶	۸۹	۹۵	۹۵	۹۷,۳	۹۷
۴	علوم‌تربیتی-مکانیک	۹۶	۹۸	۹۴	۹۶	۹۸	۹۹,۶	۹۸,۵
۵	علوم تربیتی-مدیریت	۸۹	۸۸	۸۶	۸۸	۸۷	۹۱,۲	۹۱,۲
۶	برق-مکانیک-عمران	۶۶	۴۸	۶۰	۶۶	۷۲	۸۰,۵	۷۳,۵
۷	علوم تربیتی-مدیریت-الهیات	۸۶	۸۷	۸۳	۸۸	۸۵,۵	۹۱,۷	۹۱,۷
۸	مدیریت-علوم اجتماعی - برق - مکانیک	۷۰	۷۵	۷۱	۷۴	۸۱	۸۸	۸۸
	میانگین	۸۴	۸۳	۸۰	۸۵	۸۷	۹۱	۹۰



شکل ٤-١٣تاثیر روش‌های بازنمایی کلمات در دسته‌بندی با استفاده از موضوع‌بندی کا-میانه-دوبخشی

برای بررسی‌های بیشتر با استفاده از الگوریتم خوشه‌بندی کانوپی که یکی از الگوریتم‌های خوشه‌بندی است استفاده شده است. در جدول ۴-۲۳ میزان خلوص (درست خوشه‌بندی کردن برحسب درصد) اسناد با استفاده از خوشه‌بند Canopy با استفاده از هر یک از شکل‌های نمایش کلمات نشان داده شده است.

جدول ٤-۲۳ میزان خلوص (درست خوشه‌بندی کردن برحسب درصد) اسناد با استفاده از خوشه‌بند Canopy با استفاده از هر یک از شکل‌های نمایش کلمات

ردیف	نام مجموعه‌داده مورد ارزیابی	DF	Proposed	Co-TF (min)	Co-TF (max)	LDA	DTM	DTM (tf*idf)
۱	برق-عمران	۵۱	۴۷	۴۷	۵۱	۴۶	۴۶	۴۶
۲	مدیریت – برق	۷۱	۵۰	۵۵	۵۹	۲۰	۵۳	۵۳
۳	مدیریت-عمران	۸۶	۵۶	۵۴	۸۴	۵۷	۵۵	۵۵
۴	علوم‌تربیتی-مکانیک	۴۴	۵۶	۴۹	۷۳	۵۵	۶۰	۶۰
۵	علوم تربیتی-مدیریت	۵۲	۴۹	۵۱	۴۹	۳۳	۵۹	۵۹

۶	برق-مکانیک-عمران	۳۹	۳۶	۴۰	۴۷	۲۵	۳۷	۳۷
۷	علوم تربیتی-مدیریت-الهیات	۴۵	۴۰	۴۶	۴۹	۴۷	۴۷	۴۷
۸	مدیریت-علوم اجتماعی - برق - مکانیک	۳۶	۳۹	۳۷	۴۰	۳۲	۳۲	۳۲

۴-۴-۱- محاسبه خلوص خوشه‌بندی با تعداد موضوع‌های مختلف

در این بخش با هدف بررسی عملکرد بیشتر طرح‌های پیشنهادی، برخی آزمایش‌ها را با جزئیات بیشتر در تعداد حالات مختلف در انتخاب موضوع و پارامتر فازی‌بودن الگوریتم FCM موردبررسی و ارزیابی قرار می‌دهیم. مجموعه‌داده‌های استفاده‌شده شامل رشته‌های مکانیک-علوم تربیتی، مدیریت-علوم‌تربیتی، سه رشته مهندسی و چهار رشته مدیریت-علوم اجتماعی - برق – مکانیک هستند. همان‌طور که در بخش قبل نیز به آن اشاره شد الگوریتم‌های مورداستفاده شامل موارد زیر است:

- خوشه‌بندی کا-میانه
- خوشه‌بندی فازی
- LDA

الگوریتم‌های مورداستفاده در تشکیل ماتریس‌های هم‌رخدادی علاوه بر موارد ارائه‌شده در بخش قبل روش Vec۲Word است. مطالعات نشان داده‌شده است که اندازه ابعاد برای نمایش کلمات ۵۰ تا ۱۰۰ برای این روش استفاده‌شده است (Li et al. ۲۰۱۸, Wang et al. ۲۰۱۵, Lilleberg, Zhu, and Zhang ۲۰۱۵) ما در این پژوهش از اندازه ابعاد ۷۵ استفاده کرده‌ایم. مقاله (Wang et al. ۲۰۱۵) نشان داده که مقدار پنجره ۴ برای مجموعه اسناد با اندازه کم مناسب است که اندازه پنجره نیز برابر ۴ قرار داده‌ایم. برای آموزش الگوریتم V۲W از نمونه برداری منفی نیز استفاده شده است.

آزمایش اول این بخش مربوط به مجموعه‌داده دو رشته مکانیک و علوم تربیتی است (که در دو زمینه مختلف علوم انسانی و مهندسی قرار دارند).

تحقیقات مختلفی برای انتخاب میزان فازی‌بودن[1] الگوریتم FCM انجام‌شده است. در آزمایش‌های مختلف مقدارهای متفاوتی از آن ارائه‌شده است. در (Pal and Bezdek ۱۹۹۵) مقداری بین ۱/۵ تا ۲/۵ پیشنهادشده و در (Xie and Beni ۱۹۹۱) مقدار بین ۱/۱ تا ۷ پیشنهادشده است. برای این آزمایش‌ها ما مقدار فازی‌بودن را بین ۱/۲ تا ۲/۴مورد آزمایش قرار خواهیم داد و نتایج را موردبررسی و مطالعه قرار می‌دهیم.

جدول ٤-٢٤میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۰ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۸۶.۵	۶۲.۳۱	۶۷.۲۲	۶۳.۱۶۹	۶۱.۷۶	۸۶.۵۰	۵۲	۵۵
Co-TF Min	۷۰.۶۱	۶۱.۸۸	۶۳.۰۷	۶۱.۶۵	۶۱.۶۸	۶۲.۰۸۳۵۶	۵۲	۵۵
DF	۸۱.۲۹	۶۰.۴۸	۶۴.۳۹	۶۱.۳۱	۶۱.۷۱	۷۹.۰۵	۵۲	۵۵
Proposed	۹۰.۱۷	۸۸.۲۲	۸۹.۴۳	۶۲.۳۱	۹۱.۸۳	۹۰.۷۱	۵۲	۵۵

جدول ٤-٢٥میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۳ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۵۹.۱	۶۲.۲۴	۶۶.۸۵	۶۲.۰۰	۶۱.۸۳	۹۴.۳۸	۵۱	۵۳
Co-TF Min	۶۰.۲۶	۶۱.۵۴	۶۱.۳۳	۶۱.۵۴	۶۶.۰۴	۷۰.۳۸	۵۱	۵۳
DF	۸۶.۹۶	۶۴.۳۱	۶۵.۳۷	۶۵.۶۲	۶۱.۴۰	۶۰.۶۰	۵۱	۵۳
Proposed	۸۸.۵۸	۸۳.۹۴۰	۸۳.۳۲	۸۸.۵۱	۶۱.۷۳	۹۴.۸۵	۵۱	۵۳

[1] Fuzziness (m)

جدول ٤-٢٦میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ١٦ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	١/٢	١/٥	١/٨	٢/١	٢/٤	کا-میانه	LDA	W٢V
CO-TF Max	٩١.١٣	٥٩.١	٥٩.١٥٧	٦١.٧١	٦٢.٢٤	٨٣.٤٥	٥١	٥٣
Co-TF Min	٦١.٠٢	٦٢.٧٣	٦١.٨٢	٦٢.٣٦	٦٢.٣٧	٧٠.٠٦	٥١	٥٣
DF	٨٣.٠١	٥٩.٩٣	٦٣.١٦	٦٧.١٣	٦٢.٥٤	٦٩.٩٦	٥١	٥٣
Proposed	٨٨.٠٣٦	٨٤.٩٩	٦٥.٢٧	٦١.٨٦	٦١.٧٣	٩١.٧١	٥٤	٤٥

جدول ٤-٢٧میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ١٩ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	١/٢	١/٥	١/٨	٢/١	٢/٤	کا-میانه	LDA	W٢V
CO-TF Max	٥٩.١	٦٤.٠٣	٥٩.٢٢	٦٢.٥٨	٦١.٧٠	٨٢.٧٢	٥١	٥٣
Co-TF Min	٥٩.١	٦٢.٦٣	٦١.٨١	٦٢.٤١	٦٢.٢٧	٨٥.٣٣	٥١	٥٣
DF	٨٥.٢١	٦٨.٨٨	٥٩.١	٦١.٨٥	٦١.٨١	٧٧.٣٨	٥١	٥٣
Proposed	٩١.٧٥	٨٨.٤٩	٧٨.٧٠	٦٥.٤٥	٦٢.٩١	٧٩.٨١	٥١	٥٣

جدول ٤-٢٨میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ٢٢ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	١/٢	١/٥	١/٨	٢/١	٢/٤	کا-میانه	LDA	W٢V
CO-TF Max	٦٣.٣٥	٥٩.١٥	٥٩.١	٦٣.٢٤	٧٠.١٤	٨٩.٠٦	٥٥	٤٦
Co-TF Min	٥٩.١	٧٠.١٢	٦٨.٩٦	٦٧.٧٥	٦٩.٧٣	٨٥.٩٣	٥٥	٤٦
DF	٦٦.٢٤	٥٩.١٥	٥٩.١	٦٥.٣٩	٦٨.٣٦	٨١.١٤	٥٥	٤٦
Proposed	٩١.٥٦	٨٠.٥١	٧٥.٦	٦٦.٣٠	٦٥.٨٧	٨٨.٣٧	٥٥	٤٦

جدول ٤-٢٩میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۲۵ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۸۵.۷۵	۵۹.۱	۶۱.۷۳	۶۸.۶۹	۶۹.۴۰	۶۴.۴۱	۵۵	۴۳
Co-TF Min	۶۱.۱۳	۷۲.۳۲	۶۷.۷۲	۷۲.۴۲	۶۸.۵	۸۶.۳۴	۵۵	۴۳
DF	۸۴.۵۸	۶۰.۷۱	۶۴.۳۴	۶۹.۸۶	۵۹.۳۹	۷۰.۱	۵۵	۴۳
Proposed	۸۸.۲۰	۸۴.۷۴	۶۷.۴۵	۶۶.۱۲	۶۹.۵۹	۸۳.۳۴	۵۵	۴۳

نتایج فوق نشان داده‌اند که طرح‌های ارائه‌شده در تشکیل ماتریس هم‌رخدادی و نیز استفاده از روش FCM با میزان مقدار فازی‌بودن ۱,۲ در تعداد خوشه‌های متفاوت ازنظر خلوص عملکرد بهتری را در خوشه‌بندی و میزان تفکیک‌پذیری اسناد به‌صورت غیر نظارت‌شده دارد. البته لازم به ذکر است که الگوریتم کا-میانه با الگوریتم FCM با میزان =m۲.۱ قابل‌رقابت و مقادیر خلوص آن‌ها نزدیک به یکدیگر هستند.

در ادامه میزان خلوص در ترکیب رشته‌های مدیریت-علوم‌تربیتی که موضوع‌ها توسط الگوریتم‌های FCM، K-Means، LDA، و V۲W و در حالات مختلف در تعداد موضوع‌ها و میزان فازینس بودن محاسبه‌شده، به‌دست‌آمده و در جداول آورده شده است. دو رشته انتخاب‌شده در یک حوزه یکسان علوم انسانی هستند.

جدول ٤-۳۰ میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۰ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۷۵.۴۳	۷۱.۴۳	۶۲.۳۵	۰.۵۸	۵۹.۷۵	۸۴.۸۲	٥٨	٤٠
Co-TF Min	۶۸.۴۵	۵۸.۶۰	۵۸.۶۹	۰.۵۸	۵۸.۷۶	۶۴.۹۰	٥٨	٤٠
DF	۷۶.۲۷	۶۷.۴۷	۵۸.۲۵	۰.۵۴	۵۸.۱۱	۶۶.۵۶	٥٨	٤٠
Proposed	۷۸.۴۴	۵۸.۳۲	۶۰.۲۴	۰.۵۸	۵۹.۸۵	۷۱.۷۰	٥٨	٤٠

جدول ٤-٣١میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۳ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۷۸.۶۵	۷۰.۰۶	۵۸.۱۱	۶۳.۵۱	۶۳.۳۸	۶۷.۴۶	۵۶.۰۰	۳۶.۰۰
Co-TF Min	۶۵.۸۹	۶۰.۲۶	۵۸.۱۱	۵۸.۱۲	۵۹.۶۶	۵۸.۱۱	۵۶.۰۰	۳۶.۰۰
DF	۷۵.۱۹	۶۷.۶۷	۵۸.۲۳	۵۸.۱۱	۵۸.۱۴	۸۲.۹۰	۵۶.۰۰	۳۶.۰۰
Proposed	۸۰.۰۶	۶۲.۶۳	۶۱.۸۶	۵۸.۱۱	۶۱.۴۲	۷۵.۷۸	۵۶.۰۰	۳۶.۰۰

جدول ٤-٣٢میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ١٦ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۷۵.۶۸	۰.۷۳	۶۲.۳۴	۶۲.۰۷	۶۱.۱۰	۸۰.۰۷	۶۱.۰۰	۳۹.۰۰
Co-TF Min	۶۷.۹۵	۰.۵۸	۵۸.۹۲	۵۸.۲۵	۵۹.۵۲	۶۳.۱۸	۶۱.۰۰	۳۹.۰۰
DF	۷۴.۱۱	۰.۶۲	۵۸.۲۲	۵۸.۱۲	۵۸.۱۳	۶۲.۳۴	۶۱.۰۰	۳۹.۰۰
Proposed	۷۹.۴۲	۰.۵۸	۵۹.۱۹	۵۹.۸۰	۵۸.۱۱	۷۹.۱۱	۶۱.۰۰	۳۹.۰۰

جدول ٤-٣٣میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۹ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۸۵.۲۱	۶۷.۹۶	۵۸.۲۰	۵۹.۷۹	۶۰.۵۴	۵۸.۱۱	۶۵.۰۰	۳۵.۰۰
Co-TF Min	۶۲.۴۰	۵۸.۱۱	۵۹.۹۱	۵۹.۳۰	۶۱.۱۷	۷۰.۹۲	۶۵.۰۰	۳۵.۰۰
DF	۷۱.۷۹	۵۹.۶۰	۵۸.۱۵	۵۸.۲۰	۶۱.۱۲	۷۹.۵۰	۶۵.۰۰	۳۵.۰۰
Proposed	۷۴.۸۵	۵۸.۱۸	۶۰.۶۳	۵۸.۱۱	۵۸.۲۱	۷۰.۱۷	۶۵.۰۰	۳۵.۰۰

جدول ٤-٣٤میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۲۲ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۷۵.۷۶	۶۶.۲۶	۶۳.۶۹	۶۴.۳۶	۵۹.۶۳	۶۶.۵۲	۵۷.۰۰	۳۸.۰۰
Co-TF Min	۶۶.۸۵	۶۱.۳۰	۶۰.۱۹	۵۸.۱۱	۵۸.۲۶	۶۸.۲۱	۵۷.۰۰	۳۸.۰۰
DF	۶۸.۹۰	۶۷.۵۲	۵۹.۹۸	۵۸.۱۶	۵۸.۱۱	۷۳.۲۳	۵۷.۰۰	۳۸.۰۰
Proposed	۷۱.۸۲	۶۰.۴۵	۶۱.۵۵	۶۰.۵۲	۶۱.۰۶	۶۲.۳۴	۵۷.۰۰	۳۸.۰۰

جدول ٤-۳۵میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۲۵ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۷۶.۵۴	۶۷.۷۰	۵۹.۷۳	۵۸.۶۷	۶۲.۳۰	۶۹.۰۱	۵۲.۰۰	۴۶.۰۰
Co-TF Min	۶۴.۹۱	۵۹.۱۵	۶۲.۸۹	۵۹.۱۴	۵۹.۰۴	۶۸.۵۶	۵۲.۰۰	۴۶.۰۰
DF	۵۹.۸۰	۶۸.۰۷	۶۴.۶۱	۵۸.۷۷	۵۹.۸۱	۸۱.۵۵	۵۲.۰۰	۴۶.۰۰
Proposed	۷۰.۴۱	۶۰.۳۹	۵۸.۳۸	۵۸.۱۱	۵۹.۰۶	۸۰.۳۹	۵۲.۰۰	۴۶.۰۰

برای بررسی‌های بیشتر ترکیب ۴ رشته از دو حوزه مهندسی (برق و مکانیک) و علوم انسانی (مدیریت و علوم‌تربیتی) را مورد ارزیابی قرار داده‌ایم. معیار خلوص را با استفاده از خوشه‌بند کا-میانه که موضوع‌ها توسط الگوریتم‌های FCM، K-Means، LDA، و V۲W و در حالات مختلف در تعداد موضوع‌ها و میزان فازینس بودن محاسبه‌شده، به‌دست‌آمده و در جداول آورده شده است.

جدول ٤-۳٦میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۰ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۶۴.۶۷	۵۰.۳۰	۴۵.۴۰	۴۳.۹۳	۰.۴۵	۵۱.۴۶	۷۲.۰۰	۳۳.۰۰

Co-TF Min	۴۴.۴۱	۳۸.۸۲	۳۸.۳۹	۴۰.۴۲	۰.۳۱	۴۷.۶۰	۷۲.۰۰	۳۳.۰۰
DF	۶۲.۸۱	۴۷.۲۲	۳۵.۲۹	۳۸.۰۸	۰.۴۱	۴۷.۱۲	۷۲.۰۰	۳۳.۰۰
Proposed	۶۶.۵۰	۵۱.۰۱	۵۰.۰۴	۵۰.۹۸	۰.۳۱	۶۳.۰۵	۷۲.۰۰	۳۳.۰۰

جدول ٤-۳۷میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۳ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۶۴.۶۲	۴۷.۹۲	۴۶.۶۶	۴۲.۸۵	۴۴.۲۸	۵۹.۷۸	۴۲.۰۰	۳۳.۰۰
Co-TF Min	۴۸.۲۶	۴۲.۵۸	۴۰.۰۴	۳۹.۹۷	۳۸.۲۸	۴۴.۹۶	۴۲.۰۰	۳۳.۰۰
DF	۵۵.۸۷	۴۶.۶۳	۳۸.۴۱	۴۳.۵۹	۴۱.۷۸	۴۹.۶۸	۴۲.۰۰	۳۳.۰۰
Proposed	۵۴.۶۷	۵۰.۰۰	۵۰.۵۷	۵۰.۸۳	۴۸.۳۷	۶۵.۰۸	۴۲.۰۰	۳۳.۰۰

جدول ٤-۳۸میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱٦ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۵۸.۱۲	۴۹.۶۰	۴۶.۲۱	۴۲.۴۸	۴۵.۷۳	۵۶.۲۶	۴۸.۰۰	۳۵.۰۰
Co-TF Min	۴۶.۵۲	۴۲.۸۷	۳۹.۲۰	۳۹.۷۰	۳۸.۱۲	۴۸.۷۱	۴۸.۰۰	۳۵.۰۰
DF	۴۹.۶۶	۴۸.۰۷	۴۳.۵۸	۴۳.۲۲	۴۳.۹۰	۴۸.۹۹	۴۸.۰۰	۳۵.۰۰
Proposed	۵۵.۹۰	۴۹.۷۵	۴۹.۱۹	۴۹.۳۷	۵۰.۵۴	۵۳.۳۴	۴۸.۰۰	۳۵.۰۰

جدول ٤-۳۹میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۱۹ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۵۷.۱۱	۵۰.۱۳	۴۵.۹۹	۰.۴۶	۰.۴۳	۵۰.۹۱	۵۲.۰۰	۴۴.۰۰
Co-TF Min	۴۸.۴۷	۳۴.۶۵	۳۸.۶۱	۰.۳۴	۰.۳۳	۵۲.۵۸	۵۲.۰۰	۴۴.۰۰
DF	۵۵.۵۶	۳۸.۷۵	۴۳.۷۹	۰.۴۳	۰.۴۲	۵۶.۱۳	۵۲.۰۰	۴۴.۰۰

Proposed	۵۷.۴۹	۴۵.۰۵	۴۷.۴۳	۰.۴۹	۰.۴۸	۵۲.۲۱	۵۲.۰۰	۴۴.۰۰

جدول ٤-٤٠میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۲۲ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۶۶.۱۶	۵۶.۷۶	۴۴.۳۴	۳۶.۶۵	۴۶.۵۸	۵۵.۵۷	۳۹.۰۰	۳۰.۰۰
Co-TF Min	۳۹.۷۵	۴۵.۴۶	۳۸.۱۸	۳۹.۷۸	۴۱.۹۱	۴۰.۶۲	۳۹.۰۰	۳۰.۰۰
DF	۵۵.۶۶	۶۰.۰۶	۴۶.۵۹	۳۹.۳۵	۴۲.۶۲	۴۶.۴۹	۳۹.۰۰	۳۰.۰۰
Proposed	۵۷.۴۳	۵۶.۸۸	۴۷.۳۰	۵۰.۰۲	۵۱.۸۰	۵۱.۴۰	۳۹.۰۰	۳۰.۰۰

جدول ٤-٤١میزان خلوص خوشه‌بندی با استفاده از روش‌های نمایش پیشنهادی کلمات با تعداد ۲۵ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۵۶.۷۶	۴۴.۳۴	۳۶.۶۵	۴۶.۵۸	۴۳.۶۴	۵۴.۱۱	۴۹.۰۰	۳۳.۰۰
Co-TF Min	۴۵.۴۶	۳۸.۱۸	۳۹.۷۸	۴۱.۹۱	۴۰.۹۲	۵۱.۲۸	۴۹.۰۰	۳۳.۰۰
DF	۶۰.۰۶	۴۶.۵۹	۳۹.۳۵	۴۲.۶۲	۴۴.۴۲	۵۲.۵۰	۴۹.۰۰	۳۳.۰۰
Proposed	۵۶.۸۸	۴۷.۳۰	۵۰.۰۲	۵۱.۸۰	۵۰.۸۳	۵۵.۴۲	۴۹.۰۰	۳۳.۰۰

نتایج فوق نشان داده‌اند که طرح‌های ارائه‌شده در تشکیل ماتریس هم‌رخدادی و نیز استفاده از روش FCM با میزان مقدار فازی‌بودن ۱,۲ در تعداد خوشه‌های متفاوت ازنظر خلوص عملکرد بهتری را در خوشه‌بندی و میزان تفکیک‌پذیری اسناد به‌صورت غیر نظارت‌شده دارد. البته لازم به ذکر است که الگوریتم کا-میانه با الگوریتم FCM با میزان =۲.۱m قابل‌رقابت و مقادیر خلوص آن‌ها نزدیک به یکدیگر هستند. بنابراین می‌توان از این دو روش برای موضوع‌بندی و خوشه‌بندی اسناد به‌خوبی استفاده کرد.

جدول ٤-٤۲ میانگین نتایج بدست آمده از خلوص خوشه‌بندی‌های این زیر بخش

خوشه‌بندی	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V

میانگین خلوص	۷۳	۵۸	۵۸	۶۰	۵۴	۷۱	۷۰	۵۰

۴-۴-۲- بررسی میزان دقت الگوریتم‌های پیشنهادی در دسته‌بندی رشته‌ها

به‌منظور بررسی بیشتر در خصوص تأثیر موضوع‌های تعیین‌شده در میزان دسته‌بندی رشته‌ها، دسته‌بندی مجموعه‌داده‌های مورد استفاده‌شده در بخش قبل با استفاده از الگوریتم‌های بیان‌شده مورداستفاده قرار می‌گیرند. الگوریتم مورداستفاده در دسته‌بندی الگوریتم بیزین است. پارامترهای زیادی برای ارزیابی میزان دسته‌بندی مورداستفاده قرار می‌گیرند. ازجمله این پارامترهای می‌توان به بازیابی[۱]، دقت[۲] و F-۱ (صحت) اشاره کرد[۳]. این معیارهای ارزیابی در فصل دوم موردبررسی قرارگرفته‌اند. از مهم‌ترین معیارهایی که در بازیابی اطلاعات مورداستفاده قرار می‌گیرد معیار F-۱ است که ترکیبی از معیارهای بازیابی و دقت است. در این بخش این پارامتر را در ارزیابی دسته‌بندی روش‌های پیشنهادی مورداستفاده قرار داده‌ایم. آزمایش‌ها به ترتیب بر روی مجموعه‌داده‌های مکانیک-علوم تربیتی (مهندسی و انسانی)، مدیریت-علوم تربیتی (علوم انسانی و علوم انسانی) و چهار رشته از ترکیب رشته‌های مهندسی و انسانی که در بخش قبل معرفی شد انجام‌شده است.

جدول ٤-٤٣ معیار F-1 در دسته‌بندی به روش بیزین با ۱۰ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۵	۰.۸۵	۰.۸۶	۰.۸۵	۰.۶۲	۰.۹۵	۹۷.۵۰	۷۵.۵۰
Co-TH Min	۰.۸۶	۰.۶۲	۰.۶۳	۰.۶۲	۰.۶۲	۰.۸۴	۹۷.۵۰	۷۵.۵۰
DF	۰.۹۳	۰.۸۳	۰.۷۹	۰.۷۶	۰.۷۳	۰.۹۰	۹۷.۵۰	۷۵.۵۰
Proposed	۰.۹۷	۰.۹۴	۰.۹۵	۰.۹۴	۰.۹۵	۰.۹۷	۹۷.۵۰	۷۵.۵۰

[۱] Recall

[۲] Precision

[۳] دلیل انتخاب فوق الذکر استاندارد بودن آنهاد در ارزیابی روش‌های هوشمند است. برای اطلاعات بیشتر می‌توان به کتابهای Machine Learning رجوع کرد.

جدول ٤-٤٤ معیار F-1 در دسته‌بندی به روش بیزین با ۱۳ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۶	۰.۸۴	۰.۷۹	۰.۶۲	۰.۶۲	۰.۹۵	۹۸.۶۰	۷۷.۰۰
Co-TH Min	۰.۸۸	۰.۵۰	۰.۶۱	۰.۶۲	۰.۸۱	۰.۸۹	۹۸.۶۰	۷۷.۰۰
DF	۰.۹۴	۰.۸۰	۰.۷۹	۰.۸۰	۰.۶۱	۰.۸۹	۹۸.۶۰	۷۷.۰۰
Proposed	۰.۹۷	۰.۹۴	۰.۹۴	۰.۹۵	۰.۹۴	۰.۹۷	۹۸.۶۰	۷۷.۰۰

جدول ٤-٤٥ پ معیار F-1 در دسته‌بندی به روش بیزین با ١٦ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۷	۰.۸۴	۰.۸۱	۰.۶۲	۰.۶۲	۰.۹۶	۹۸.۲۰	۷۷.۰۰
Co-TH Min	۰.۸۲	۰.۴۸	۰.۶۲	۰.۶۲	۰.۶۲	۰.۸۷	۹۸.۲۰	۷۷.۰۰
DF	۰.۹۴	۰.۸۳	۰.۸۰	۰.۷۹	۰.۷۹	۰.۹۱	۹۸.۲۰	۷۷.۰۰
Proposed	۰.۹۷	۰.۹۵	۰.۹۴	۰.۹۴	۰.۹۳	۰.۹۸	۹۸.۲۰	۷۷.۰۰

جدول ٤-٤٦ معیار F-1 در دسته‌بندی به روش بیزین با ۱۹ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۶	۰.۸۹	۰.۸۰	۰.۷۰	۰.۷۶	۰.۹۳	۹۸.۰۰	۸۰.۰۰
Co-TH Min	۰.۷۴	۰.۷۴	۰.۷۴	۰.۷۳	۰.۷۳	۰.۹۳	۹۸.۰۰	۸۰.۰۰
DF	۰.۹۳	۰.۸۷	۰.۸۰	۰.۷۸	۰.۷۹	۰.۹۲	۹۸.۰۰	۸۰.۰۰
Proposed	۰.۹۷	۰.۹۴	۰.۹۴	۰.۹۳	۰.۹۱	۰.۹۵	۹۸.۰۰	۸۰.۰۰

جدول ٤-٤٧ معیار F-1 در دسته‌بندی به روش بیزین با ۲۲ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۵	۰.۸۴	۰.۸۳	۰.۸۰	۰.۷۰	۰.۹۵	۹۷.۸۰	۸۰.۰۰
Co-TH Min	۰.۷۶	۰.۷۰	۰.۶۹	۰.۶۸	۰.۷۰	۰.۹۵	۹۷.۸۰	۸۰.۰۰
DF	۰.۹۱	۰.۸۳	۰.۸۱	۰.۸۰	۰.۶۸	۰.۹۳	۹۷.۸۰	۸۰.۰۰
Proposed	۰.۹۷	۰.۹۴	۰.۹۵	۰.۹۶	۰.۹۱	۰.۹۵	۹۷.۸۰	۸۰.۰۰

جدول ٤-٤٨ معیار F-1 در دسته‌بندی به روش بیزین با ۲۵ موضوع برای رشته‌های مکانیک و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۹۶	۰.۷۹	۰.۸۸	۰.۶۹	۰.۶۹	۰.۹۶	۹۷.۵۰	۸۰.۰۰
Co-TH Min	۰.۷۳	۰.۷۴	۰.۶۸	۰.۷۲	۰.۶۹	۰.۹۶	۹۷.۵۰	۸۰.۰۰
DF	۰.۹۳	۰.۸۵	۰.۸۱	۰.۷۸	۰.۷۹	۰.۹۲	۹۷.۵۰	۸۰.۰۰
Proposed	۰.۹۶	۰.۹۴	۰.۹۴	۰.۹۴	۰.۷۰	۰.۹۷	۹۷.۵۰	۸۰.۰۰

در ادامه معیار F-۱ را با استفاده از دسته‌بند بیزین در دسته‌بندی رشته‌های مدیریت–علوم‌تربیتی توسط الگوریتم‌های FCM، K-Means، LDA، و V۲W و در حالات مختلف در تعداد موضوع‌ها و میزان فازینس بودن محاسبه‌شده و در جداول آورده شده است. دو رشته انتخاب‌شده در یک حوزه یکسان علوم انسانی هستند.

جدول ٤-٤٩ معیار F-1 در دسته‌بندی به روش بیزین با ۱۰ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۸۹	۰.۷۵	۰.۵۸	۶۱.۱۰	۰.۵۸	۰.۸۸	۸۳.۰۰	۶۴.۰۰

Co-TH Min	۰.۷۵	۰.۵۸	۰.۵۸	۵۹.۸۶	۰.۵۸	۰.۸۲	۸۳.۰۰	۶۴.۰۰
DF	۰.۸۶	۰.۷۱	۰.۶۰	۵۸.۲۲	۰.۵۸	۰.۸۲	۸۳.۰۰	۶۴.۰۰
Proposed	۰.۸۸	۰.۵۸	۰.۵۸	۶۰.۱۰	۰.۵۸	۰.۸۶	۸۳.۰۰	۶۴.۰۰

جدول ٤-٥٠ معیار F-1 در دسته‌بندی به روش بیزین با ۱۳ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	½	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۰.۸۹	۰.۷۱	۰.۶۰	۰.۵۸	۰.۵۸	۰.۸۶	۸۱.۷۰	۶۵.۰۰
Co-TH Min	۰.۷۴	۰.۵۸	۰.۵۸	۰.۵۸	۰.۵۸	۰.۷۹	۸۱.۷۰	۶۵.۰۰
DF	۰.۸۷	۰.۶۷	۰.۵۵	۰.۵۸	۰.۵۸	۰.۸۹	۸۱.۷۰	۶۵.۰۰
Proposed	۰.۸۹	۰.۵۸	۰.۵۸	۰.۵۸	۰.۵۸	۰.۸۹	۸۱.۷۰	۶۵.۰۰

جدول ٤-٥١ معیار F-1 در دسته‌بندی به روش بیزین با ١٦ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۰.۸۹	۷۰.۲۸	۰.۵۸	۰.۵۸	۰.۵۸	۰.۸۸	۸۴.۰۰	۶۹.۰۰
Co-TH Min	۰.۷۲	۶۱.۶۴	۰.۵۸	۰.۵۸	۰.۵۸	۰.۸۳	۸۴.۰۰	۶۹.۰۰
DF	۰.۸۵	۵۸.۱۱	۰.۵۶	۰.۵۸	۰.۵۸	۰.۸۴	۸۴.۰۰	۶۹.۰۰
Proposed	۰.۸۸	۵۸.۵۲	۰.۵۸	۰.۵۸	۰.۵۸	۰.۸۷	۸۴.۰۰	۶۹.۰۰

جدول ٤-٥٢ معیار F-1 در دسته‌بندی به روش بیزین با ۱۹ موضوع برای رشته‌های مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W2V
CO-TF Max	۰.۸۹	۰.۷۴	۰.۶۴	۰.۵۴	۰.۵۷	۰.۸۸	۸۵.۰۰	۷۳.۰۰
Co-TH Min	۰.۶۷	۰.۴۹	۰.۵۵	۰.۵۱	۰.۵۶	۰.۸۴	۸۵.۰۰	۷۳.۰۰
DF	۰.۸۷	۰.۶۵	۰.۵۵	۰.۵۳	۰.۵۷	۰.۸۷	۸۵.۰۰	۷۳.۰۰

W2V	LDA	کا-میانه	۲/٤	۲/۱	۱/۸	۱/٥	۱/۲	M
۷۳.۰۰	۸٥.۰۰	۰.۸۶	۰.٥۸	۰.٥۷	۰.۶۰	۰.٥۹	۰.۸۸	Proposed

جدول ٤-٥٣ معیار F-1 در دسته‌بندی به روش بیزین با ۲۲ موضوع برای رشته‌های مدیریت و علوم تربیتی

W2V	LDA	کا-میانه	۲/٤	۲/۱	۱/۸	۱/٥	۱/۲	M
۷٥.۰۰	۸٥.۶۰	۰.۸۶	۰.٥۸	۰.٥۸	۰.٥۸	۰.۷۳	۰.۹۰	CO-TF Max
۷٥.۰۰	۸٥.۶۰	۰.۸٥	۰.٥۸	۰.٥۸	۰.٥۹	۰.٥۹	۰.۶۹	Co-TH Min
۷٥.۰۰	۸٥.۶۰	۰.۸۷	۰.٥۸	۰.٥۸	۰.۶۲	۰.۶۹	۰.۸۳	DF
۷٥.۰۰	۸٥.۶۰	۰.۸۶	۰.٥۸	۰.٥۸	۰.٥۷	۰.٥۸	۰.۸۸	Proposed

جدول ٤-٥٤ معیار F-1 در دسته‌بندی به روش بیزین با ۲٥ موضوع برای رشته‌های مدیریت و علوم تربیتی

W2V	LDA	کا-میانه	۲/٤	۲/۱	۱/۸	۱/٥	۱/۲	M
۷۱.۶۰	۸۷.۰۰	۰.۸۸	۰.۶۰	۰.٥۸	۰.۶۰	۰.۷۲	۰.۸۸	CO-TF Max
۷۱.۶۰	۸۷.۰۰	۰.۸۴	۰.٥۹	۰.٥۸	۰.٥۹	۰.٥۹	۰.۷۰	Co-TH Min
۷۱.۶۰	۸۷.۰۰	۰.۸۹	۰.۶۲	۰.۶۴	۰.۶۸	۰.۷۶	۰.۸٥	DF
۷۱.۶۰	۸۷.۰۰	۰.۸۷	۰.٥۸	۰.٥۸	۰.٥۸	۰.٥۹	۰.۸۹	Proposed

برای بررسی‌های بیشتر ترکیب ۴ رشته از دو حوزه مهندسی (برق و مکانیک) و علوم انسانی (مدیریت و علوم‌تربیتی) را مورد ارزیابی قرار داده‌ایم. معیار F-۱ را با استفاده از دسته‌بند بیزین توسط الگوریتم‌های FCM، K-Means، LDA، و V۲W و در حالات مختلف در تعداد موضوع‌ها و میزان فازینس بودن محاسبه‌شده و در جداول آورده شده است.

جدول ٤-٥٥ معیار F-1 در دسته‌بندی به روش بیزین با ۱۰ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

W2V	LDA	کا-میانه	۲/٤	۲/۱	۱/۸	۱/٥	۱/۲	M
٥۰.۰۰	۸۰.۰۰	۰.۶۹	۴۴.۰۲	۰.۴۸	۰.۴۸	۰.٥۴	۰.۷۳	CO-TF Max

Co-TH Min	۰.۵۸	۰.۴۶	۰.۴۳	۰.۳۱	۳۹.۰۹	۰.۶۰	۸۰.۰۰	۵۰.۰۰
DF	۰.۷۳	۰.۵۱	۰.۴۷	۰.۴۶	۴۱.۷۳	۰.۶۹	۸۰.۰۰	۵۰.۰۰
Proposed	۰.۷۴	۰.۵۳	۰.۳۲	۰.۳۱	۵۰.۹۳	۰.۷۲	۸۰.۰۰	۵۰.۰۰

جدول ٤-٥٦ معیار F-1 در دسته‌بندی به روش بیزین با ۱۳ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱/۲	۱/۵	۱/۸	۲/۱	۲/٤	کا-میانه	LDA	W۲V
CO-TF Max	۰.۷۶	۰.۵۲	۰.۴۸	۰.۴۷	۰.۴۴	۰.۷۲	۸۰.۰۰	۵۲.۰۰
Co-TH Min	۰.۶۲	۰.۴۵	۰.۳۷	۰.۳۲	۰.۳۲	۰.۶۵	۸۰.۰۰	۵۲.۰۰
DF	۰.۷۱	۰.۴۸	۰.۴۶	۰.۴۲	۰.۴۱	۰.۷۰	۸۰.۰۰	۵۲.۰۰
Proposed	۰.۷۴	۰.۵۱	۰.۳۱	۰.۳۱	۰.۳۱	۰.۷۷	۸۰.۰۰	۵۲.۰۰

جدول ٤-٥٧ معیار F-1 در دسته‌بندی به روش بیزین با ١٦ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱.۲۰	۱.۵۰	۱.۸۰	۲.۱۰	۲.٤۰	KMEANS	LDA	W۲V
CO-TF Max	۰.۷۶	۰.۵۰	۰.۴۶	۰.۴۶	۰.۴۵	۰.۷۲	۸۰.۰۰	۵۰.۰۰
Co-TH Min	۰.۶۱	۰.۴۳	۰.۳۹	۰.۳۱	۰.۳۲	۰.۶۵	۸۰.۰۰	۵۰.۰۰
DF	۰.۷۱	۰.۵۲	۰.۴۵	۰.۴۲	۰.۴۲	۰.۷۱	۸۰.۰۰	۵۰.۰۰
Proposed	۰.۷۵	۰.۵۲	۰.۳۱	۰.۳۱	۰.۳۲	۰.۷۲	۸۰.۰۰	۵۰.۰۰

جدول ٤-٥٨ معیار F-1 در دسته‌بندی به روش بیزین با ۱۹ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱.۲۰	۱.۵۰	۱.۸۰	۲.۱۰	۲.٤۰	KMEANS	LDA	W۲V
CO-TF Max	۰.۷۸	۰.۵۸	۰.۴۷	۴۵.۲۳	۴۳.۸۲	۰.۷۴	۸۲.۰۰	۵۳.۷۰
Co-TH Min	۰.۶۰	۰.۴۴	۰.۳۸	۳۶.۷۱	۳۹.۲۲	۰.۷۱	۸۲.۰۰	۵۳.۷۰
DF	۰.۷۲	۰.۵۰	۰.۴۷	۴۳.۷۰	۴۲.۸۲	۰.۷۴	۸۲.۰۰	۵۳.۷۰

Proposed	۰.۷۶	۰.۴۹	۰.۴۶	۴۸.۸۸	۵۱.۱۹	۰.۷۴	۸۲.۰۰	۵۳.۷۰

جدول ٤-٥٩ معیار F-1 در دسته‌بندی به روش بیزین با ۲۲ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱.۲۰	۱.۵۰	۱.۸۰	۲.۱۰	۲.٤۰	KMEANS	LDA	W۲V
CO-TF Max	۰.۷۶	۰.۷۷	۰.۵۶	۰.۵۲	۰.۴۹	۰.۷۸	۸۱.۵۰	۵۰.۰۰
Co-TH Min	۰.۶۳	۰.۵۹	۰.۴۶	۰.۳۴	۰.۳۴	۰.۶۹	۸۱.۵۰	۵۰.۰۰
DF	۰.۷۳	۰.۷۱	۰.۵۰	۰.۴۸	۰.۴۱	۰.۷۴	۸۱.۵۰	۵۰.۰۰
Proposed	۰.۷۹	۰.۷۷	۰.۵۳	۰.۳۳	۰.۳۳	۰.۸۰	۸۲.۰۰	۵۳.۰۰

جدول ٤-٦٠ معیار F-1 در دسته‌بندی به روش بیزین با ۲۵ موضوع برای رشته‌های برق مکانیک مدیریت و علوم تربیتی

M	۱.۲۰	۱.۵۰	۱.۸۰	۲.۱۰	۲.٤۰	KMEANS	LDA	W۲V
CO-TF Max	۰.۷۷	۰.۵۶	۰.۵۲	۰.۴۹	۰.۴۳	۰.۷۵	۸۱.۵۰	۵۰.۰۰
Co-TH Min	۰.۵۹	۰.۴۶	۰.۳۴	۰.۳۴	۰.۳۳	۰.۷۳	۸۱.۵۰	۵۰.۰۰
DF	۰.۷۱	۰.۵۰	۰.۴۸	۰.۴۱	۰.۴۴	۰.۷۵	۸۱.۵۰	۵۰.۰۰
Proposed	۰.۷۷	۰.۵۳	۰.۳۳	۰.۳۳	۰.۳۳	۰.۷۹	۸۱.۵۰	۵۰.۰۰

نتایج فوق نشان داده‌اند که طرح‌های ارائه‌شده در تشکیل ماتریس هم‌رخدادی و نیز استفاده از روش FCM با میزان مقدار فازی‌بودن ۱,۲ در تعداد خوشه‌های متفاوت ازنظر خلوص عملکرد بهتری را در خوشه‌بندی و میزان تفکیک‌پذیری اسناد به‌صورت غیر نظارت‌شده دارد. البته لازم به ذکر است که الگوریتم کا-میانه با الگوریتم FCM با میزان =m۲.۱ قابل‌رقابت و مقادیر خلوص آن‌ها نزدیک به یکدیگر هستند. بنابراین می‌توان از این دو روش برای موضوع‌بندی و خوشه‌بندی اسناد به‌خوبی استفاده کرد.

۴-۴-۳- تعیین تعداد موضوع‌های یک حوزه

به‌منظور انتخاب بهینه تعداد موضوع‌های یک حوزه علمی، در این بخش روشی استفاده‌شده (Peng, Zhang, Kou, and Shi ۲۰۱۲Peng, Zhang, Kou, Li, et al. ۲۰۱۲,) که با به‌کارگیری از معیارهای ارزیابی داخلی و روش تصمیم‌گیری چند مقدارِ تعداد بهینه‌ای از موضوع‌ها ازنظر ساختار داخلی (ویژگی‌های ذاتی) آن‌ها پیشنهاد می‌شود. در این رساله از معیارهای درونی سیلوهت[۱] و میزان فاصله از مرکز خوشه‌ها (SSE) استفاده کرده‌ایم. تعداد خوشه‌ها را از دو تا سی تغییر داده و در هر حالت شاخص‌های ارزیابی خوشه‌بندی را محاسبه کرده‌ایم. ازآنجاکه در بخش بعدی روند حوزه برق را قرار است تحلیل کنیم از این رشته برای پیاده‌سازی این طرح استفاده کرده‌ایم.

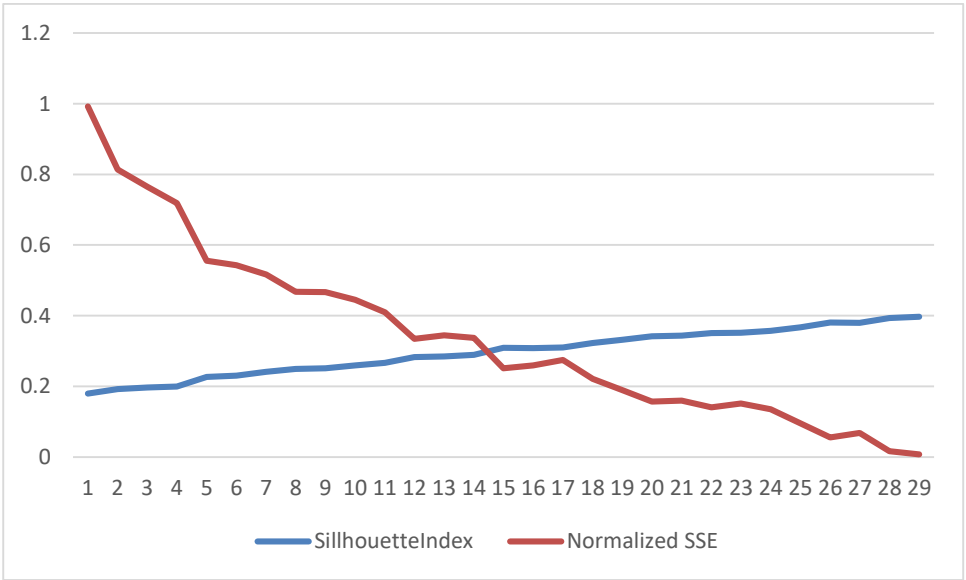
جدول ٦١٤-نتایج شاخص‌های سیلوهت و SSE برای تعداد خوشه‌های مختلف در حوزه برق

تعداد خوشه	سیلوهت	SSE
۲	۰.۱۸	۵.۳۳
۳	۰.۱۸	۴.۹۸
۴	۰.۲۰	۴.۷۸
۵	۰.۲۱	۴.۶۴
۶	۰.۲۳	۴.۳۲
۷	۰.۲۳	۴.۲۷
۸	۰.۲۴	۴.۱۶
۹	۰.۲۵	۴.۱۵
۱۰	۰.۲۶	۴.۰۵
۱۱	۰.۲۵	۴.۱۴
۱۲	۰.۲۷	۳.۹۳
۱۳	۰.۲۸	۳.۹۱
۱۴	۰.۲۹	۳.۷۶
۱۵	۰.۲۹	۳.۸۴
۱۶	۰.۳۰	۳.۷۶
۱۷	۰.۳۱	۳.۶۸

[۱] Silouhet۱۷

۱۸	۰.۳۱	۳.۶۹
۱۹	۰.۳۲	۳.۶۶
۲۰	۰.۳۳	۳.۵۴
۲۱	۰.۳۳	۳.۵۱
۲۲	۰.۳۴	۳.۴۳
۲۳	۰.۳۵	۳.۳۹
۲۴	۰.۳۴	۳.۴۶
۲۵	۰.۳۷	۳.۲۳
۲۶	۰.۳۷	۳.۲۱
۲۷	۰.۳۸	۳.۲۳
۲۸	۰.۳۹	۳.۰۹
۲۹	۰.۴۰	۳.۰۷
۳۰	۰.۴۰	۳.۰۷

نتایج جدول 614- نشان داده است که تعداد ۱۷ خوشه‌های ازنظر شاخص‌های ارزیابی تعداد خوشه بهتری نسبت به مابقی هستند اما ازآنجاکه تعداد خوشه‌های ۱۳ تا ۱۸ ازنظر عددی مشابه این تعداد هستند می‌توانند در گزینه‌های انتخابی نیز قرار گیرند.



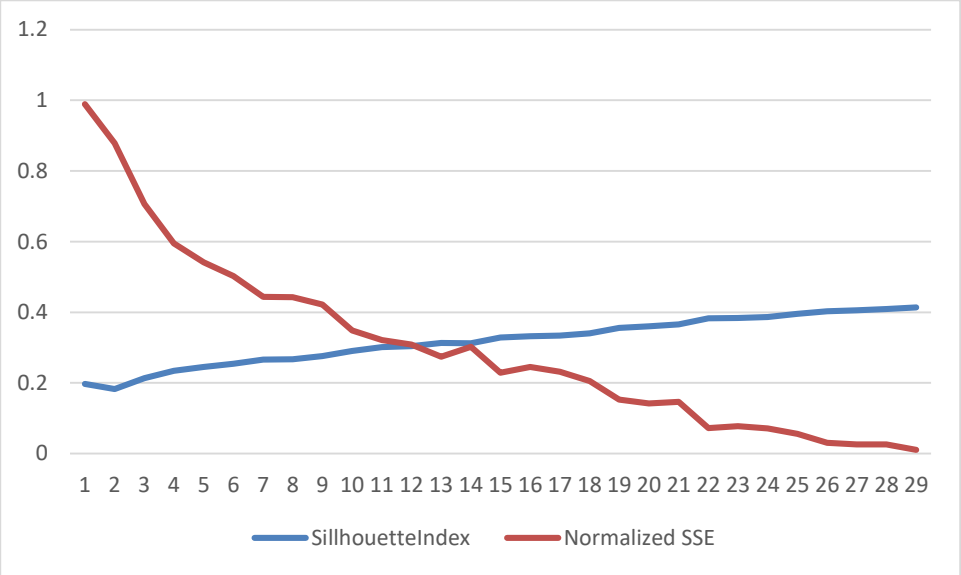
شکل ٤-١٤ مقادیر شاخص‌های مختلف به ازای تعداد خوشه‌های مختلف برای رشته مهندسی برق

برای رشته مکانیک مقادیر شاخص‌های ارزیابی خوشه‌بندی به‌قرار زیر است

جدول ٤-٦٢ شاخص‌های اندازه گرفته‌شده خوشه برای رشته مکانیک و رتبه هر تعداد خوشه

تعداد خوشه	SillhouetteIndex	Normalized SSE
۲	۰.۱۹۶۷۱۵	۰.۹۸۹
۳	۰.۱۸۲۲۶۴	۰.۸۷۸۵
۴	۰.۲۱۳۴۴۲	۰.۷۰۶۷۸۹
۵	۰.۲۳۴۳۹۲	۰.۵۹۴۸۶۴
۶	۰.۲۴۵۰۶۶	۰.۵۴۱۵۶۵
۷	۰.۲۵۴۰۱۳	۰.۵۰۲۴۷
۸	۰.۲۶۶۱۳۹	۰.۴۴۳۲۵
۹	۰.۲۶۶۸۳۸	۰.۴۴۳۱۳۷
۱۰	۰.۲۷۶۱۵۸	۰.۴۲۱۷۲۸
۱۱	۰.۲۹۰۴۴۶	۰.۳۴۸۳۰۱
۱۲	۰.۳۰۱۳۷۴	۰.۳۲۰۸۶۱
۱۳	۰.۳۰۳۸۹۲	۰.۳۰۸۹۲۸
۱۴	۰.۳۱۳۰۶۲	۰.۲۷۳۹۷۸
۱۵	۰.۳۱۲۰۴۳	۰.۳۰۲۰۹۸
۱۶	۰.۳۲۸۶۱۲	۰.۲۲۸۷۳
۱۷	۰.۳۳۲۲۸۱	۰.۲۴۵۵۱
۱۸	۰.۳۳۳۷۷۲	۰.۲۳۱۱۳۵
۱۹	۰.۳۴۰۳۰۱	۰.۲۰۵۱۷
۲۰	۰.۳۵۵۶۵۸	۰.۱۵۳۰۸۷
۲۱	۰.۳۶۰۶۱۷	۰.۱۴۲۰۰۶
۲۲	۰.۳۶۵۴۹۲	۰.۱۴۶۷۹۸
۲۳	۰.۳۸۲۸۱۱	۰.۰۷۲۴۶۷
۲۴	۰.۳۸۳۸۲۲	۰.۰۷۷۱۱۵
۲۵	۰.۳۸۶۸۸۳	۰.۰۷۱۳۵۷
۲۶	۰.۳۹۵۵۷۶	۰.۰۵۵۸۶۹
۲۷	۰.۴۰۲۳۸	۰.۰۳۰۶۲۵
۲۸	۰.۴۰۵۱۱۳	۰.۰۲۵۷۰۳

۲۹	۰.۴۰۹۲۱۴	۰.۰۲۶۰۱۶
۳۰	۰.۴۱۳۳۴۸	۰.۰۱۰۵۹۳



شکل ٤-١٥ مقادیر شاخص‌های مختلف به ازای تعداد خوشه‌های مختلف برای رشته مهندسی مکانیک

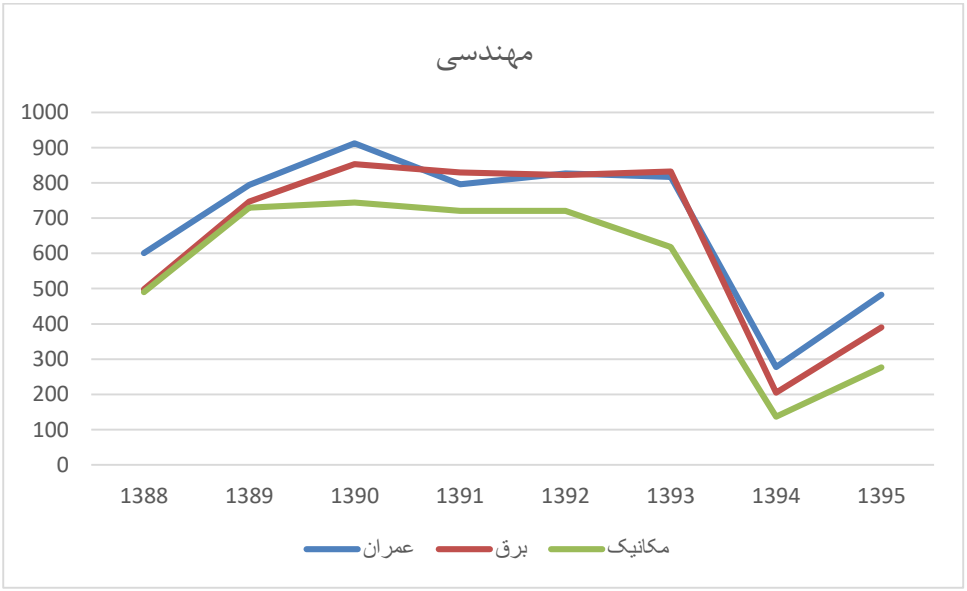
۴-۵- تحلیل روند

در ابتدا به بررسی میزان رشد اسناد استخراج‌شده از پایگاه گنج برحسب سال می‌پردازیم. همان‌طور که جدول زیر نشان می‌دهد که روند رشد رساله‌ها از سال ۱۳۸۸ تا سال ۱۳۹۲ را نشان می‌دهد.

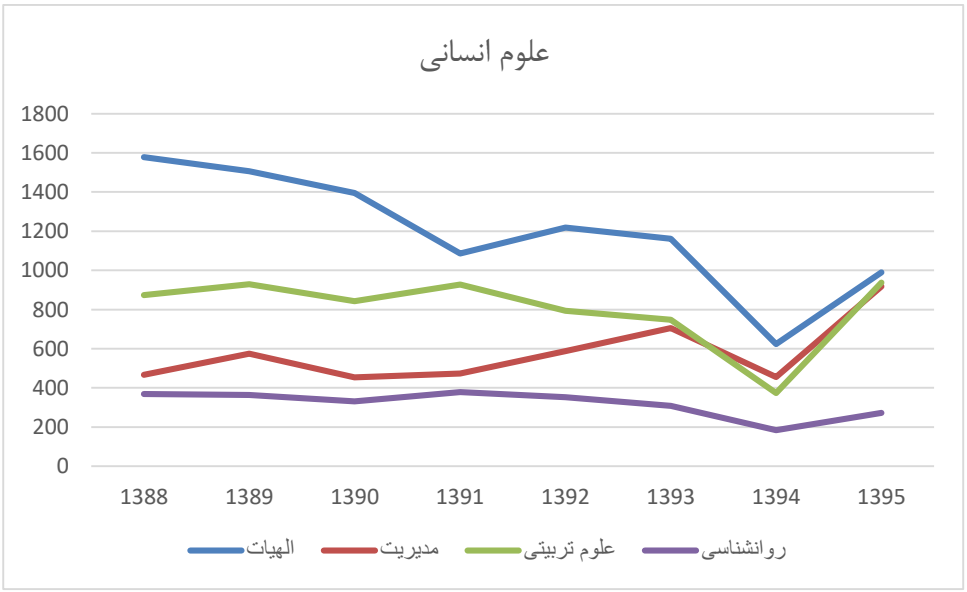
جدول ٤-٦٣ آمار پارساهای استخراج‌شده در رشته‌های مختلف در سال‌های مشخص

	الهیات	مدیریت	علوم تربیتی	روانشناسی	عمران	برق	مکانیک
۱۳۸۸	۱۵۷۸	۴۶۷	۸۷۴	۳۶۸	۶۰۱	۴۹۸	۴۹۰
۱۳۸۹	۱۵۰۶	۵۷۴	۹۳۰	۳۶۴	۷۹۴	۷۴۷	۷۳۰
۱۳۹۰	۱۳۹۵	۴۵۳	۸۴۲	۳۳۱	۹۱۲	۸۵۳	۷۴۴
۱۳۹۱	۱۰۸۶	۴۷۳	۹۲۸	۳۷۸	۷۹۶	۸۳۰	۷۲۱

۱۳۹۲	۱۲۱۸	۵۸۷	۷۹۴	۳۵۳	۸۲۷	۸۲۲	۷۲۱
۱۳۹۳	۱۱۶۱	۷۰۶	۷۴۸	۳۰۹	۸۱۷	۸۳۲	۶۱۸
۱۳۹٤	۶۲۴	۴۵۵	۳۷۴	۱۸۴	۲۷۸	۲۰۵	۱۳۷
۱۳۹۵	۹۹۰	۹۱۸	۹۳۸	۲۷۳	۴۸۳	۳۹۰	۲۷۷



شکل ٤-١٦ روند رشته‌های مهندسی در طی زمان

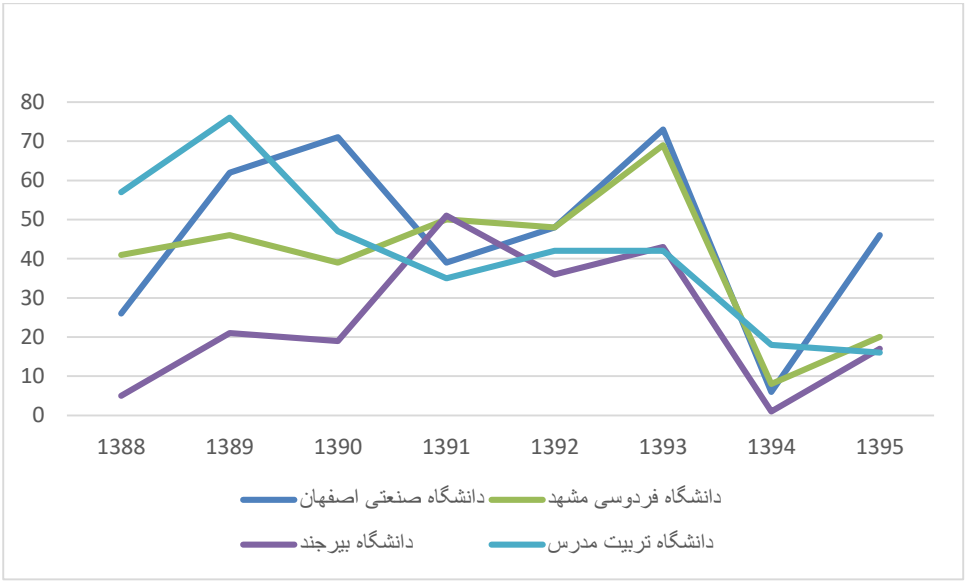


شکل ٤-١٧روند رشته‌های علوم انسانی استخراج‌شده در طی زمان

در این بخش برای نشان دادن نحوه عملکرد روش تحلیل‌روند ارائه‌شده در فصل قبل، یکی از رشته‌های استخراج‌شده از پایگاه داده گنج را مورد تحلیل و بررسی قرار خواهیم داد و روند آن را بررسی خواهیم داد. رشته انتخاب‌شده رشته برق است که درمجموع از **۷۵** دانشگاه کشور در سامانه ثبت کرده‌اند.

جدول ٤-٦٤ پژوهش‌های استخراج‌شده از سامانه گنج در حوزه برق

	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹٤	۱۳۹٥
تعداد سند	۴۹۸	۷۴۷	۸۵۳	۸۳۰	۸۲۲	۸۳۲	۲۰۵	۳۹۰



شکل ٤-١٨ برخی دانشگاه‌هایی که بیشترین پژوهش‌ها را در حوزه برق در سامانه ثبت کرده‌اند

شکل ۴-۱۸ برخی دانشگاه‌هایی که بیشترین پژوهش‌ها را در حوزه برق در سامانه ثبت کرده‌اند را نشان می‌دهد. این شکل نشان می‌دهد که ثبت پارساهای دانشگاه‌ها در بازه زمانی ۱۳۹۴ و ۱۳۹۵ نسبت به سال‌های قبل افت شدیدی کرده است. این افت می‌تواند ناشی از سیاست‌ها کلان کشوری و وزارتخانه مربوطه باشد.

در بخش قبل دیدیم که بر اساس پارامترهای درون خوشه‌ای تعداد دوازده موضوع (خوشه) برای رشته برق انتخاب شد کلمات برتر هر موضوع به‌قرار زیر است.

جدول ٤-٦٥ مشخصات موضوع‌های تشکیل‌شده در رشته برق

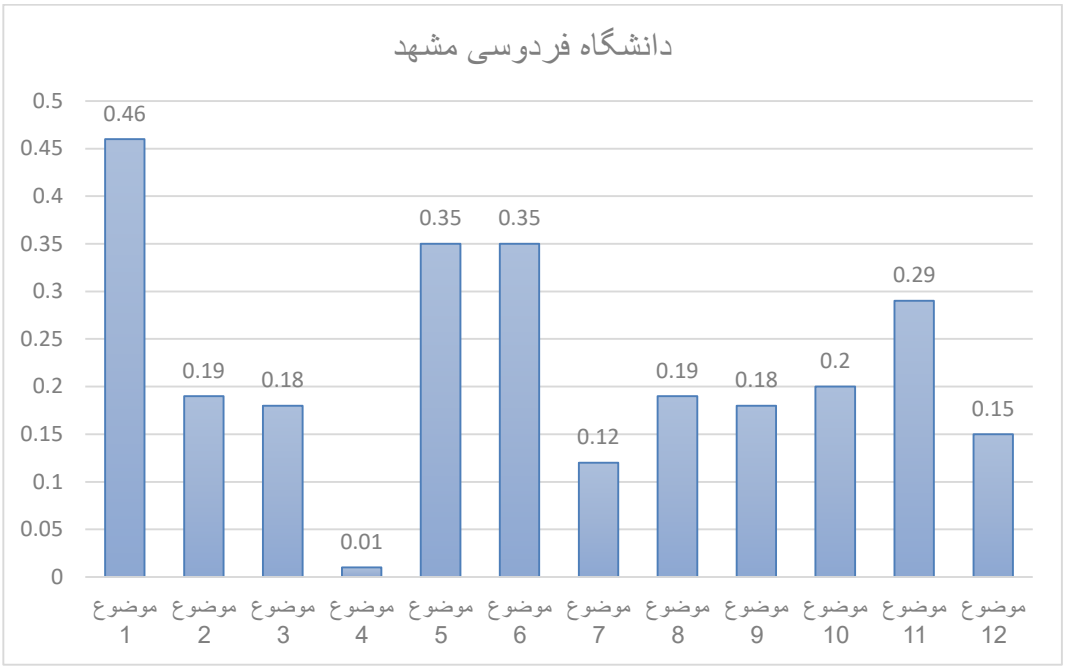
	مجموع تکرار کلمات	مجموع تعداد اسناد به ازای کلمات	تعداد کلمات	کلمات	تعداد استفاده موضوع‌ها در پارساها	نرمال‌شده هر موضوع در هر پارسا
موضوع ۱	۱۶۴۳۲	۴۳۹۵	۲۰	شبکه، بهبود، بررسی، تحلیل بهینه‌سازی، الگوریتم، کاهش، مدل‌سازی، جریان	۲۸۱۷	۰/۵۴
موضوع ۲	۱۴۱۹	۱۳۳۷	۱۴	شبکه_حسگر_بی‌سیم، تبدیل_موجک پردازش_تصویر، استخراج_ویژگی، پهنای_نوار، پردازش سیگنال	۱۱۷۱	۰/۲۳
موضوع ۳	۶۶۷۴	۹۸۴	۱۰	توربین، ترانسفورماتور، نیروگاه، ژنراتور، بازار، ریزشبکه، پایداری، حفاظت، مبدل	۸۴۰	۰/۱۶
موضوع ۴	۴۱۶	۷۹	۱	تبدیل	۷۹	۰/۰۲
موضوع ۵	۱۸۵۳	۱۱۵۱	۱۴	تطبیقی، شناسایی، سیگنال، مدل فیلتر، تخمین، فازی، خطا، تشخیص، آنالیز	۱۴۴۳	۰/۲۸
موضوع ۶	۳۵۷۷	۳۲۲۸	۲۴	منطق_فازی، انرژی_الکتریکی، کنترل_ولتاژ، توزیع_نیروی_برق، کنترل_فازی، شبیه‌سازی_رایانه‌ای، کیفیت_توان، شبکه_توزیع، الگوریتم_ژنتیک،	۲۰۰۳	۰/۳۹

			قابلیت_اطمینان، توربین_بادی، سیستم_قدرت، تولید_پراکنده، شبکه_عصبی			
موضوع ۷	۳۵۱۷	۵۷۰	۶	فرکانسی، سنکرون، حذف، سرعت، زمان، کانال	۵۳۵	۰/۱۰
موضوع ۸	۷۱۷۴	۱۱۱۹	۱۱	سیگنال، پردازش، آنالوگ، مخابرات، آرایه، ردیابی، آشکارسازی، نوری، تصویر، آنتن	۹۳۶	۰/۱۸
موضوع ۹	۸۳۹۲	۱۲۶۲	۱۱	تکنولوژی، CMOS، مدار، خازن، فاز، نویز، موتور، حسگر، باند	۹۸۵	۰/۱۹
موضوع ۱۰	۴۶۱۴	۱۰۴۵	۷	مدلسازی، حضور، ارائه، اثر، ساختار، فرکانس	۹۳۰	۰/۱۸
موضوع ۱۱	۱۲۲۲۹	۲۲۳۰	۱۸	توسعه، ظرفیت، خطوط، مدیریت، اطمینان، برنامه‌ریزی، تلفات، برق، قابلیت، انتقال، منابع، انرژی، تولید	۱۳۴۷	۰/۲۶
موضوع ۱۲	۸۰۸	۵۹۲	۷	شبیه‌سازی_سیستم_قدرت، بهینه_سازی_گرده_ذرات، واحدهای نیروگاه_بادی، عدم_قطعیت، بازار_برق، تأثیر	۴۹۱	۰/۰۹

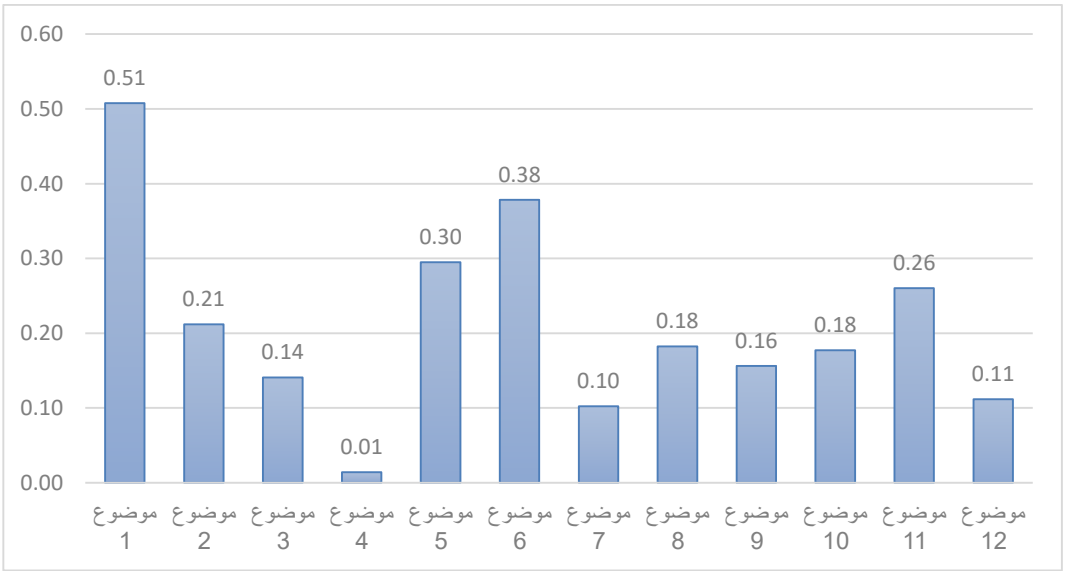
درمجموع ۵۱۷۷ پارسا، این ۱۲ موضوع ۱۳۵۷۵ بار در اسناد تکرار شده که به‌صورت متوسط هر پارسا ۲/۶ موضوع مورداستفاده قرارگرفته است. پرطرفدارترین و جذاب‌ترین موضوع‌های استفاده‌شده برای محققان همان‌طور که در جدول ۴-۶۵ می‌بینید موضوع‌های ۱، ۶، ۵ و ۱۱ هستند. لازم به ذکر است که تنها برای سه درصد از پارساهای این حوزه موضوعی پیدا نشد که می‌توان آن‌ها را جزو داده‌های پرت محسوب کرد.

مجدول ٤-٦٦ متوسط سهم موضوع‌های انجام‌شده در یک رساله در دانشگاه‌های با بیشترین پارساهای ثبت‌شده (برای مقایسه پذیری اعداد داخل جدول برحسب تعداد پارساها نرمال شده‌اند و در ۱۰۰ ضرب شده)

موضوع / نام دانشگاه	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲	میانگین تعداد موضوع در یک پارسا
دانشگاه صنعتی خواجه نصیرالدین طوسی	۰.۵۸	۰.۲۱	۰.۱۹	۰.۰۱	۰.۲۷	۰.۲۹	۰.۱۲	۰.۱۷	۰.۲۰	۰.۱۸	۰.۲۴	۰.۰۹	۲.۵۴
دانشگاه تبریز	۰.۵۱	۰.۲۹	۰.۱۲	۰.۰۳	۰.۲۸	۰.۳۲	۰.۱۰	۰.۲۰	۰.۲۰	۰.۱۵	۰.۲۱	۰.۰۴	۲.۴۵
دانشگاه صنعتی اصفهان	۰.۶۲	۰.۲۲	۰.۱۲	۰.۰۱	۰.۲۷	۰.۲۶	۰.۱۲	۰.۲۰	۰.۱۴	۰.۱۴	۰.۲۳	۰.۰۷	۲.۴۱
دانشگاه تربیت مدرس	۰.۴۶	۰.۱۳	۰.۱۸	۰.۰۱	۰.۲۳	۰.۳۱	۰.۰۸	۰.۱۳	۰.۱۴	۰.۱۹	۰.۲۸	۰.۱۱	۲.۲۵
دانشگاه فردوسی مشهد	۰.۴۶	۰.۱۹	۰.۱۸	۰.۰۱	۰.۳۵	۰.۳۵	۰.۱۲	۰.۱۹	۰.۱۸	۰.۲۰	۰.۲۹	۰.۱۵	۲.۶۷
دانشگاه صنعتی شاهرود	۰.۳۸	۰.۲۴	۰.۰۹	۰.۰۱	۰.۴۵	۰.۵۲	۰.۰۹	۰.۱۹	۰.۱۸	۰.۱۴	۰.۱۷	۰.۰۸	۲.۵۴
دانشگاه بیرجند	۰.۶۱	۰.۲۷	۰.۱۳	۰.۰۳	۰.۱۸	۰.۴۷	۰.۱۰	۰.۲۰	۰.۱۳	۰.۲۲	۰.۳۴	۰.۱۶	۲.۸۳
میانگین	۰,۵۱	۰,۲۱	۰,۱۴	۰,۰۱	۰,۳۰	۰,۳۸	۰,۱۰	۰,۱۸	۰,۱۶	۰,۱۸	۰,۲۶	۰,۱۱	۲,۵۴



شکل ٤-١٩متوسط سهم استفاده هریک از موضوع‌ها در یک رساله انجام شده در دانشگاه فردوسی مشهد



شکل ٤-٢٠ متوسط سهم استفاده هریک از موضوع‌ها در یک رساله

مجدول ۴-۶۶ متوسط سهم موضوع‌های انجام‌شده در یک رساله در دانشگاه‌های با بیشترین پارساهای ثبت‌شده (برای مقایسه پذیری اعداد داخل جدول برحسب تعداد پارساها نرمال شده‌اند و در ۱۰۰ ضرب شده) را نشان می‌دهد. با استفاده از این جدول می‌توان فهمید که کدام دانشگاه‌ها بیشتر بر روی چه موضوع‌هایی تمرکز کرده‌اند. برای مثال بیشتر دانشگاه‌ها بر روی موضوع یک فعالیت می‌کنند. برای مثال دیگر دانشگاه بیرجند بیشتر از دانشگاه‌های دیگر بر روی موضوع ۱۱ کار می‌کند. دانشگاه شاهرود بر روی موضوع ۵ بیشتر و دانشگاه تبریز بر روی موضوع ۲ تمرکز بیشتری دارد.

۴-۵-۱-۱- تحلیل موضوع و نمودار راهبردی با استفاده از ماتریس هم‌رخدادی موضوعی

با استفاده از تشکیل ماتریس موضوع-سند می‌توانیم میزان مشارکت موضوع‌ها را در اسناد یک حوزه علمی مورد تجزیه تحلیل قرار دهیم. با این تحلیل می‌توان مشخص کرد که موضوع‌ها به چه میزان در کل اسناد شرکت کرده‌اند، پرطرفدارترین موضوع کدام است، ارتباط بین موضوع‌ها در اسناد به چه صورت است. برای نمایش این امر از نمودار راهبردی استفاده می‌کنیم. در نمودار راهبردی استفاده‌شده محور افقی میزان مرکزیت[۱] (که به‌عنوان میان‌رشته‌ای بودن یک موضوع نیز اشاره می‌شود) و محور عمودی میزان چگالی[۲] (که بلوغ یک موضوع نیز گفته می‌شود) هر موضوع را نشان می‌دهد. در نمودار راهبردی ارائه‌شده میزان مرکزیت میزان ارتباط بین هر موضوع با موضوع‌های دیگر به دست می‌آید و بلوغ (چگالی) هر موضوع بر اساس تعداد دفعاتی که بر روی آن موضوع در پارساها از آن‌ها استفاده‌شده است. در این رساله نمودار راهبردی به دو صورت ترسیم و موردبررسی قرارگرفته می‌شود: بر اساس ماتریس هم‌رخدادی موضوع‌ها، بر اساس ماتریس هم‌رخدادی کلمات.

بنابراین برای ترسیم نمودار راهبردی در گام اول با استفاده از ماتریس موضوع-سند، ماتریس هم‌رخدادی موضوع‌ها را تشکیل خواهیم داد. با استفاده از ماتریس هم‌رخدادی موضوع‌ها، مقادیر

---

[۱] Centrality

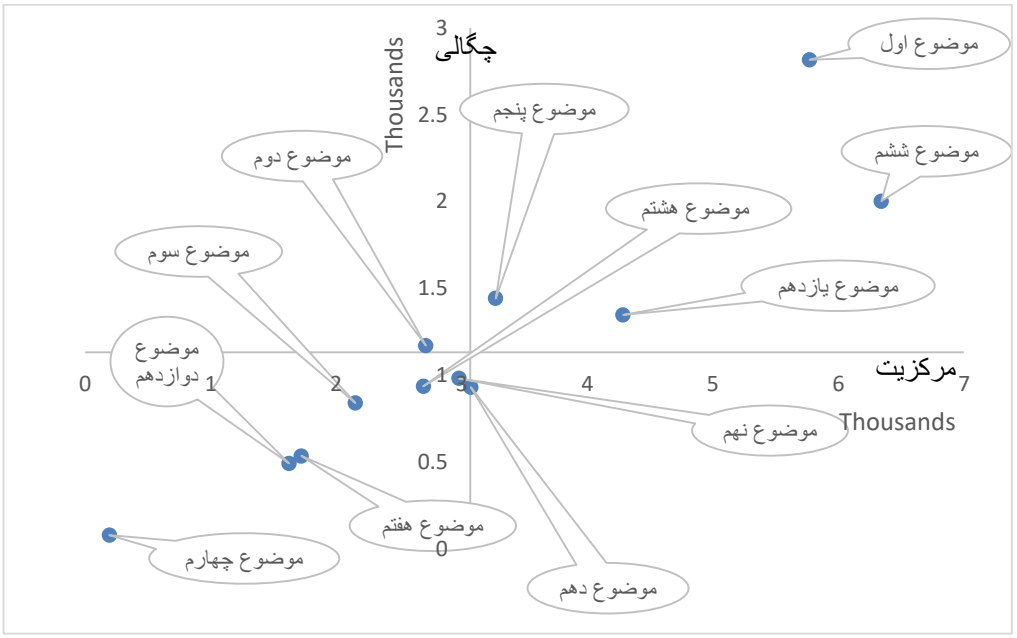
[۲] Density

مرکزیت و چگالی موضوع‌ها محاسبه‌شده و نمودار راهبردی را ترسیم می‌کنیم. در گام دوم با استفاده از ماتریس هم‌رخدادی کلمات، مقادیر مرکزیت و بلوغ محاسبه‌شده و سپس نمودار راهبردی ترسیم خواهد شد.

جدول ٤-٦٧ محاسبه مرکزیت و بلوغ موضوع‌ها بر اساس ماتریس هم‌رخدادی موضوع‌ها

نام خوشه	مرکزیت	بلوغ
موضوع ۱	۵۷۷۱	۲۸۱۷
موضوع ۲	۲۷۱۲	۱۱۷۱
موضوع ۳	۲۱۵۲	۸۴۰
موضوع ٤	۱۹۲	۷۹
موضوع ٥	۳۲۷۰	۱۴۴۳
موضوع ٦	۶۳۴۳	۲۰۰۳
موضوع ۷	۱۷۲۱	۵۳۵
موضوع ۸	۲۶۹۴	۹۳۶
موضوع ۹	۲۹۷۸	۹۸۳
موضوع ۱۰	۳۰۷۳	۹۳۰
موضوع ۱۱	۴۲۸۴	۱۳۴۷
موضوع ۱۲	۱۶۲۴	۴۹۱
میانگین	۳۰۶۷	۱۱۳۱

بر اساس جدول ۴-۶۷ با استفاده از میزان بلوغ و مرکزیت رشته‌های نمودار راهبردی آن به‌صورت می‌شود.



شکل ٤-٢١ نمودار راهبردی بر اساس ماتریس هم‌رخدادی موضوع‌ها

ازآنجاکه یال‌های گراف فوق بسیار زیاد بوده و ترسیم آن شکل را ناخوانا می‌کند، ماتریس یال‌ها را در ... آورده‌ایم. در این ماتریس یال‌هایی که قدرت آن‌ها بیشتر از ۱۵ درصد مرکزیت (ارتباط بین موضوع و موضوع‌های دیگر) دارند مشخص‌شده است.

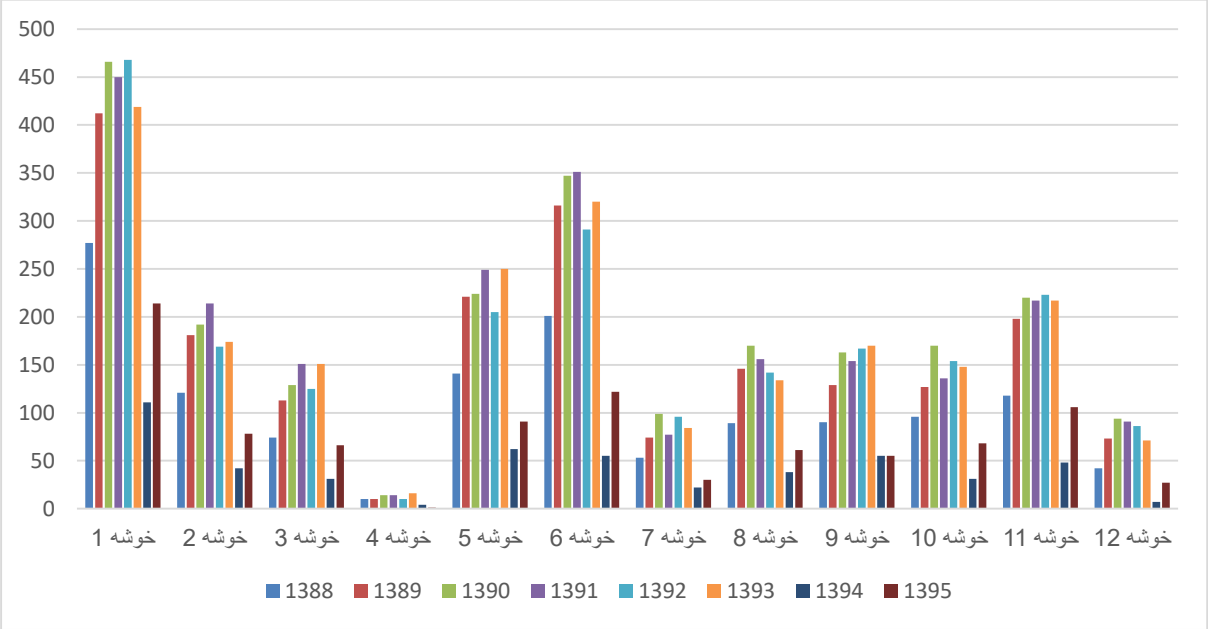
جدول ٤-٦٨ ماتریس یال‌های نمودار راهبردی

موضوع- موضوع	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲
موضوع ۱	۱	۱	۰	۰	۱	۱	۰	۰	۰	۰	۱	۰
موضوع ۲	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
موضوع ۳	۱	۰	۱	۰	۰	۱	۰	۰	۰	۰	۰	۰
موضوع ٤	۰	۱	۰	۱	۰	۰	۰	۱	۰	۰	۰	۰
موضوع ٥	۱	۰	۰	۰	۱	۱	۰	۰	۰	۰	۰	۰
موضوع ٦	۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰
موضوع ۷	۱	۰	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰
موضوع ۸	۱	۰	۰	۰	۰	۰	۰	۱	۰	۰	۰	۰
موضوع ۹	۱	۰	۰	۰	۰	۰	۰	۰	۱	۰	۰	۰
موضوع ۱۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۰
موضوع ۱۱	۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۱	۰
موضوع ۱۲	۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۱	۱

بر اساس روند ماتریس سند-موضوع‌ها، روند حرکت هر موضوع به‌صورت جدول ۴-۷۳ استخراج‌شده است.

جدول ٤-٦٩ روند فراوانی هر موضوع در سال‌های متوالی

	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹٤	۱۳۹٥
موضوع ۱	۲۷۷	۴۱۲	۴۶۶	۴۵۰	۴۶۸	۴۱۹	۱۱۱	۲۱۴
موضوع ۲	۱۲۱	۱۸۱	۱۹۲	۲۱۴	۱۶۹	۱۷۴	۴۲	۷۸
موضوع ۳	۷۴	۱۱۳	۱۲۹	۱۵۱	۱۲۵	۱۵۱	۳۱	۶۶
موضوع ٤	۱۰	۱۰	۱۴	۱۴	۱۰	۱۶	۴	۱
موضوع ٥	۱۴۱	۲۲۱	۲۲۴	۲۴۹	۲۰۵	۲۵۰	۶۲	۹۱
موضوع ٦	۲۰۱	۳۱۶	۳۴۷	۳۵۱	۲۹۱	۳۲۰	۵۵	۱۲۲
موضوع ۷	۵۳	۷۴	۹۹	۷۷	۹۶	۸۴	۲۲	۳۰
موضوع ۸	۸۹	۱۴۶	۱۷۰	۱۵۶	۱۴۲	۱۳۴	۳۸	۶۱
موضوع ۹	۹۰	۱۲۹	۱۶۳	۱۵۴	۱۶۷	۱۷۰	۵۵	۵۵
موضوع ۱۰	۹۶	۱۲۷	۱۷۰	۱۳۶	۱۵۴	۱۴۸	۳۱	۶۸
موضوع ۱۱	۱۱۸	۱۹۸	۲۲۰	۲۱۷	۲۲۳	۲۱۷	۴۸	۱۰۶
موضوع ۱۲	۴۲	۷۳	۹۴	۹۱	۸۶	۷۱	۷	۲۷



شکل ٤-٢٢ روند رشد موضوع‌ها در سال‌های متوالی بر اساس ماتریس هم‌رخدادی موضوع‌ها

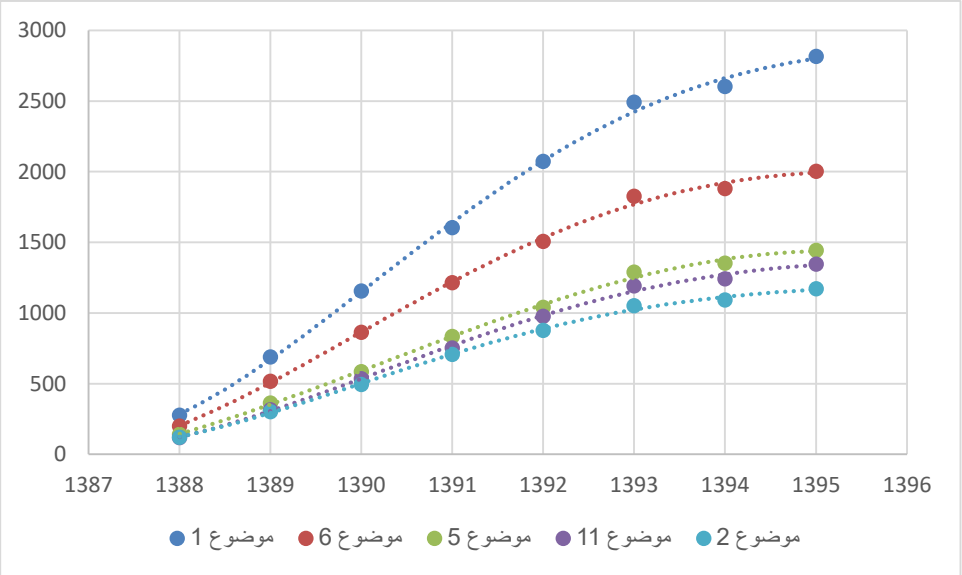
فراوانی تجمعی یک دسته یا محدوده حاصل جمع فراوانی مطلق همان دسته یا رده با فراوانی مطلق دسته یا رده های قبلی می باشد. با استفاده از فراوانی تجمعی می‌توانیم از تعداد اسنادی که در موضوعی خاصی کار کرده‌اند آگاه شویم. جدول ۴-۷۰ این فراوانی را برای موضوع‌های مختلف نشان می‌دهد.

جدول ٤-٧٠ فراوانی تجمعی اسناد انجام شده در هر موضوع در سال‌های متوالی

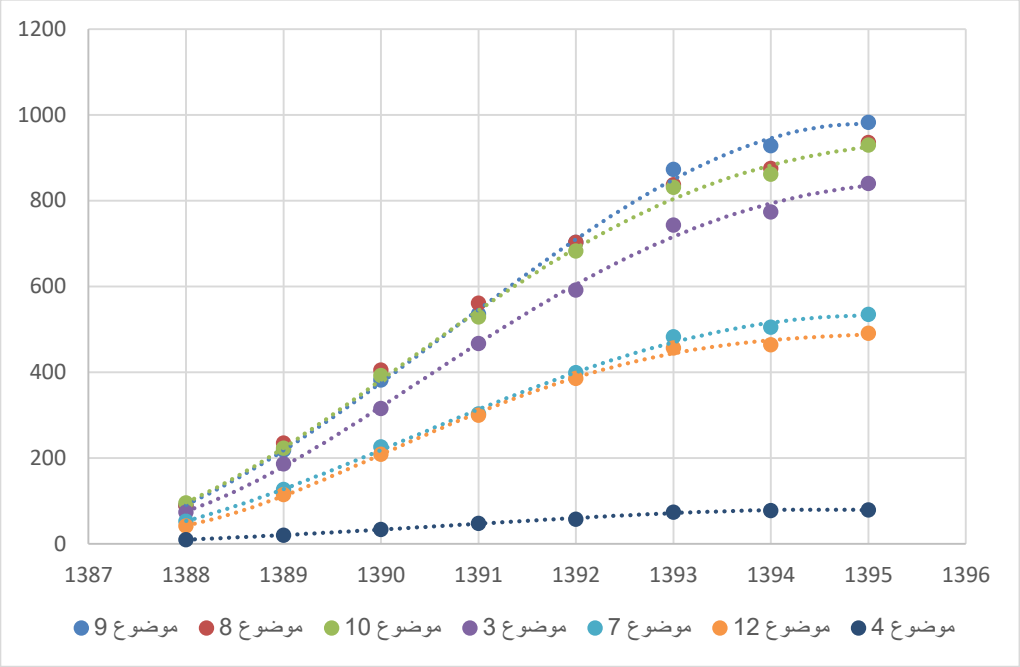
	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴	۱۳۹۵
موضوع ۱	۲۷۷	۶۸۹	۱۱۵۵	۱۶۰۵	۲۰۷۳	۲۴۹۲	۲۶۰۳	۲۸۱۷
موضوع ۲	۱۲۱	۳۰۲	۴۹۴	۷۰۸	۸۷۷	۱۰۵۱	۱۰۹۳	۱۱۷۱
موضوع ۳	۷۴	۱۸۷	۳۱۶	۴۶۷	۵۹۲	۷۴۳	۷۷۴	۸۴۰
موضوع ۴	۱۰	۲۰	۳۴	۴۸	۵۸	۷۴	۷۸	۷۹
موضوع ۵	۱۴۱	۳۶۲	۵۸۶	۸۳۵	۱۰۴۰	۱۲۹۰	۱۳۵۲	۱۴۴۳
موضوع ۶	۲۰۱	۵۱۷	۸۶۴	۱۲۱۵	۱۵۰۶	۱۸۲۶	۱۸۸۱	۲۰۰۳
موضوع ۷	۵۳	۱۲۷	۲۲۶	۳۰۳	۳۹۹	۴۸۳	۵۰۵	۵۳۵
موضوع ۸	۸۹	۲۳۵	۴۰۵	۵۶۱	۷۰۳	۸۳۷	۸۷۵	۹۳۶
موضوع ۹	۹۰	۲۱۹	۳۸۲	۵۳۶	۷۰۳	۸۷۳	۹۲۸	۹۸۳
موضوع ۱۰	۹۶	۲۲۳	۳۹۳	۵۲۹	۶۸۳	۸۳۱	۸۶۲	۹۳۰

موضوع ۱۱	۱۱۸	۳۱۶	۵۳۶	۷۵۳	۹۷۶	۱۱۹۳	۱۲۴۱	۱۳۴۷
موضوع ۱۲	۴۲	۱۱۵	۲۰۹	۳۰۰	۳۸۶	۴۵۷	۴۶۴	۴۹۱

در درصورتی که جدول فراوانی تجمعی این موضوع‌ها را رسم کنیم روند تکاملی این موضوع‌ها را نشان می‌دهد. این روند حجم اسنادی در موضوع‌ها را تا آن لحظه نشان می‌دهد. در شکل ۴-۲۳ و شکل ۴-۲۴ نمودار روند موضوع‌ها را با استفاده از نمودار چند جمله‌ای درجه ۴ روسم شده است. این نمودارها نشان می‌دهند که روند کارکرد بر روی موضوع‌ها در حال توقف است.



شکل ٤-۲۳ نمودار روند فراوانی تجمعی موضوع‌های پر کاربردتر

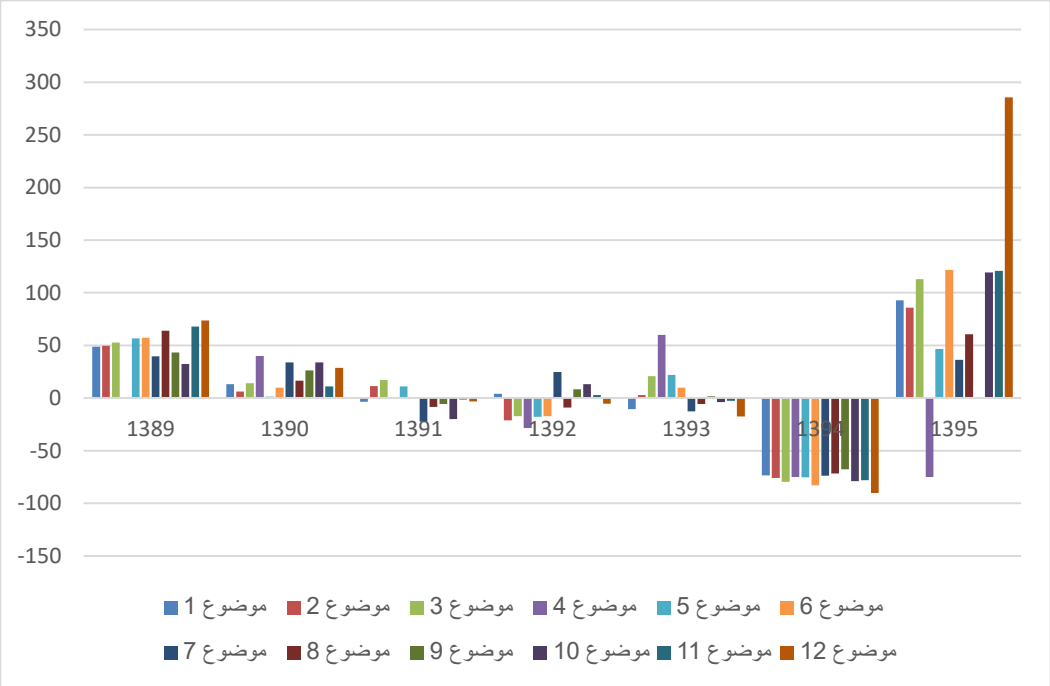


شکل ٤-٢٤ نمودار روند فراوانی تجمعی موضوع‌های با کاربرد کمتر

با استفاده از جدول ۴-۷۳ می‌توان نرخ رشد هر موضوع را محاسبه کرد. این نرخ رشد موضوع در سال‌های متوالی به‌قرار زیر است.

جدول ٤-٧١ نرخ رشد موضوع‌ها نسبت به سال گذشته برحسب درصد (هم‌رخدادی موضوعی)

	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴	۱۳۹۵
موضوع ۱	۴۹	۱۳	-۳	۴	-۱۰	-۷۴	۹۳
موضوع ۲	۵۰	۶	۱۱	-۲۱	۳	-۷۶	۸۶
موضوع ۳	۵۳	۱۴	۱۷	-۱۷	۲۱	-۷۹	۱۱۳
موضوع ٤	۰	۴۰	۰	-۲۹	۶۰	-۷۵	-۷۵
موضوع ٥	۵۷	۱	۱۱	-۱۸	۲۲	-۷۵	۴۷
موضوع ٦	۵۷	۱۰	۱	-۱۷	۱۰	-۸۳	۱۲۲
موضوع ۷	۴۰	۳۴	-۲۲	۲۵	-۱۳	-۷۴	۳۶
موضوع ۸	۶۴	۱۶	-۸	-۹	-۶	-۷۲	۶۱
موضوع ۹	۴۳	۲۶	-۶	۸	۲	-۶۸	۰
موضوع ۱۰	۳۲	۳۴	-۲۰	۱۳	-۴	-۷۹	۱۱۹
موضوع ۱۱	۶۸	۱۱	-۱	۳	-۳	-۷۸	۱۲۱
موضوع ۱۲	۷۴	۲۹	-۳	-۵	-۱۷	-۹۰	۲۸۶



شکل ٤-٢٥ نرخ رشد موضوع‌ها نسبت به سال گذشته برحسب درصد (هم‌رخدادی موضوعی)

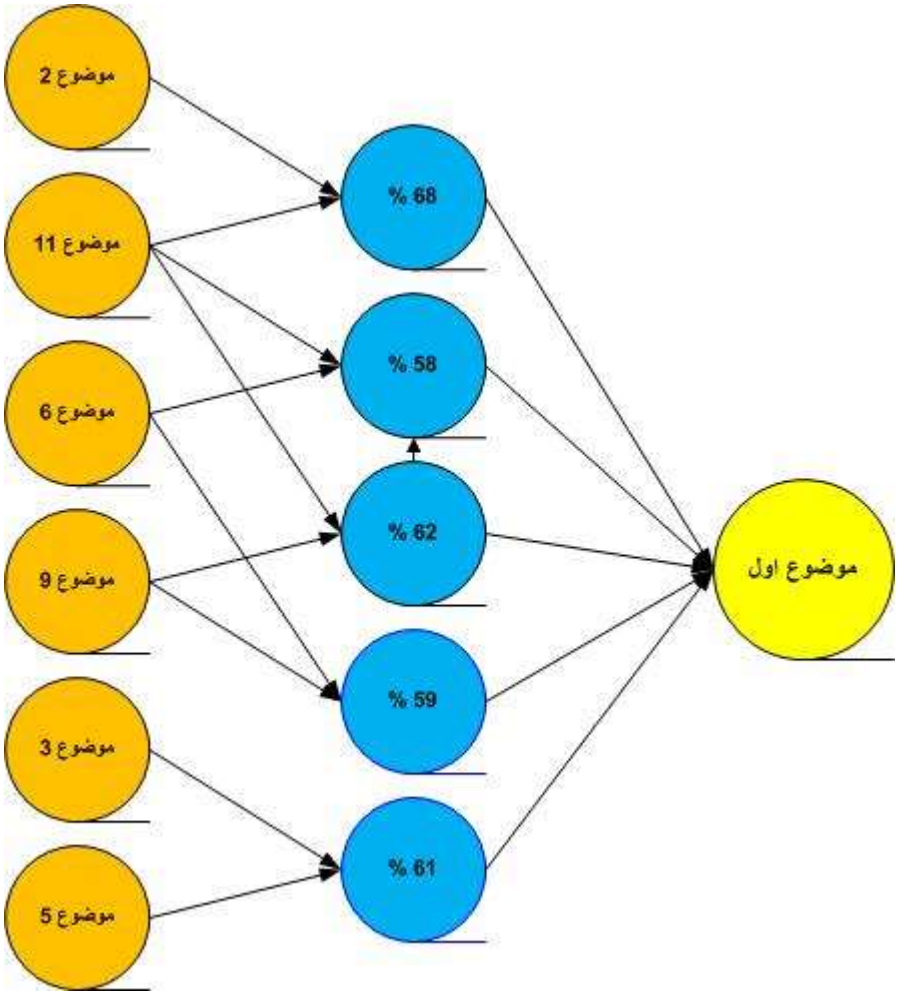
با استفاده از قوانین انجمنی می‌توان ارتباط احتمال اینکه پژوهشگری برای یک پژوهش با انتخاب یک موضوع به چه موضوع‌های دیگری می‌پردازد را محاسبه کرد. استفاده از قوانین انجمنی ازآنجا قابل اهمیت است که به خاطر کثرت تعداد پارساها در مجموعه متون و نیز ترکیب‌های مختلف موضوع‌ها استخراج این قوانین به‌صورت دستی توسط فرد متخصص امری بسیار زمان‌بر است. قوانین انجمنی استخراج‌شده در بین موضوع‌ها در جدول ۴-۷۲ برخی از قوانین انجمنی استخراج‌شده ارتباط بین موضوع‌ها نشان داده‌شده است.

جدول ٤-٧٢ برخی از قوانین انجمنی استخراج‌شده ارتباط بین موضوع‌ها

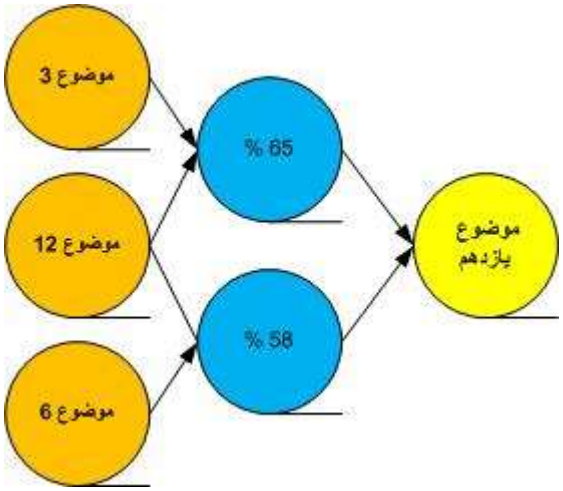
پیامد (موضوع)	شرط (موضوع)	Confidence	Suppor
۱	۱۱ ،۲	۰.۶۸	۱۲۷
۶	۱۲ ،۱۱	۰.۶۸	۱۷۷
۶	۱۱ ،۳	۰.۶۶	۲۱۲
۶	۱۱ ،۲ ،۱	۰.۶۶	۱۱۱

۱۱۰	۰.۶۵	۱۲ ،۳	۱۱
۱۲۵	۰.۶۴	۱۰ ،۳	۶
۱۰۷	۰.۶۳	۱۲ ،۳	۶
۱۸۸	۰.۶۳	۱۱ ،۱۰	۶
۱۶۶	۰.۶۲	۱۲ ،۱	۶
۳۰۴	۰.۶۲	۱۲	۶
۱۱۵	۰.۶۲	۱۱ ،۹	۱
۱۲۴	۰.۶۱	۵ ،۳	۱
۵۰۰	۰.۶۰	۳	۶
۱۶۵	۰.۵۹	۹ ،۶	۱
۴۳۵	۰.۵۸	۱۱ ،۶	۱
۱۱۹	۰.۵۸	۵ ،۳	۶
۱۷۷	۰.۵۸	۱۲ ،۶	۱۱

شکل ۴-۲۶ روابط بین موضوع‌ها که منجر به وقوع موضوع اول می‌شود با استفاده از قوانین انجمنی را نشان می‌دهد. داخل هر دایره میزان اطمنیان در صورتی که دو موضوع فعال (مورد استفاده قرار گرفته) شده باشند را نشان می‌دهد.برای مثال این شکل نشان می‌دهد که در صورتی موضوع‌های دوم و یازدهم با هم در یک سند استفاده شده باشند با احتمال و اطمنیان ۶۸ درصد موضوع اول نیز مورد توجه نویسنده قرار گرفته‌است و از آن استفاده کرده است. این روابط قدرت میان‌رشته‌ای بودن موضوع‌ها را نشان می‌دهد. از طرفی دیگر هرچقدر موضوعی در موضوع دیگری نقش داشته باشد و با هم مورد استفاده قرار گرفته باشند، نزدیکی بین دو موضوع و تاثیر آن‌ها بر روی یکدگیر را نشان می‌دهد. این امر را می‌توان با استفاده از شمارش تعداد یال‌های خروجی هر موضوع فهمید. موضوع‌هایی که در شکل‌های زیر با رنگ نارنجی نشان داده شده اند بیشترین ارتباط را با موضوع زرد رنگ از بین مابقیه موضوع‌ها دارند هرچقدر یال خروجی موضوع نارنجی رنگ بیشتر باشد آن موضوع ارتباط و تاثیر بیشتری با موضوع زرد رنگ دارد.

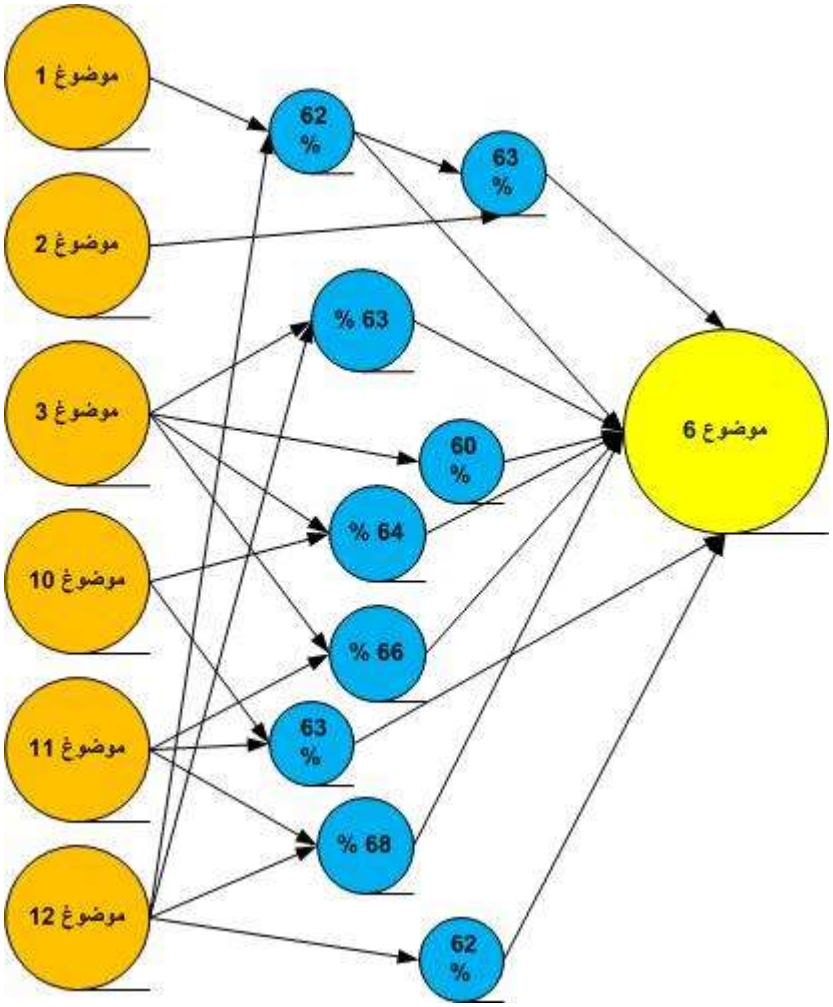


شکل ٤-٢٦ روابط بین موضوع‌ها که منجر به وقوع موضوع اول می‌شود با استفاده از قوانین انجمنی



شکل ٤-٢٧ روابط بین موضوع‌ها که منجر به وقوع موضوع یازدهم می‌شود با استفاده از قوانین انجمنی

همانطور که در شکل ۴-۲۷ ملاحظه می‌کنید در صورتی که موضوع سه و دوازده در سندی با هم مورد استفاده نویسنده قرار بگیرند احتمال اینکه موضوع دوازده نیز در آن سند مورد توجه نویسنده قرار بگیرد چیزی حدودر ۶۵ درصد است. علاوه بر آن در صورتی که موضوع ششم نیز همراه با موضوع دوازدهم مورد استفاده قرار بگیرد با احتمال ۵۸ درصد در آن سند موضوع یازدهم نیز یکی از موضوع‌های اصلی است. بدلیل اینکه یال‌های خروجی از گره شمال دوازده بیشتر از بقیه است این شکل نشان می‌دهد که موضوع دوازدهم ارتباط زیادی با موضوع یازدهم دارد.

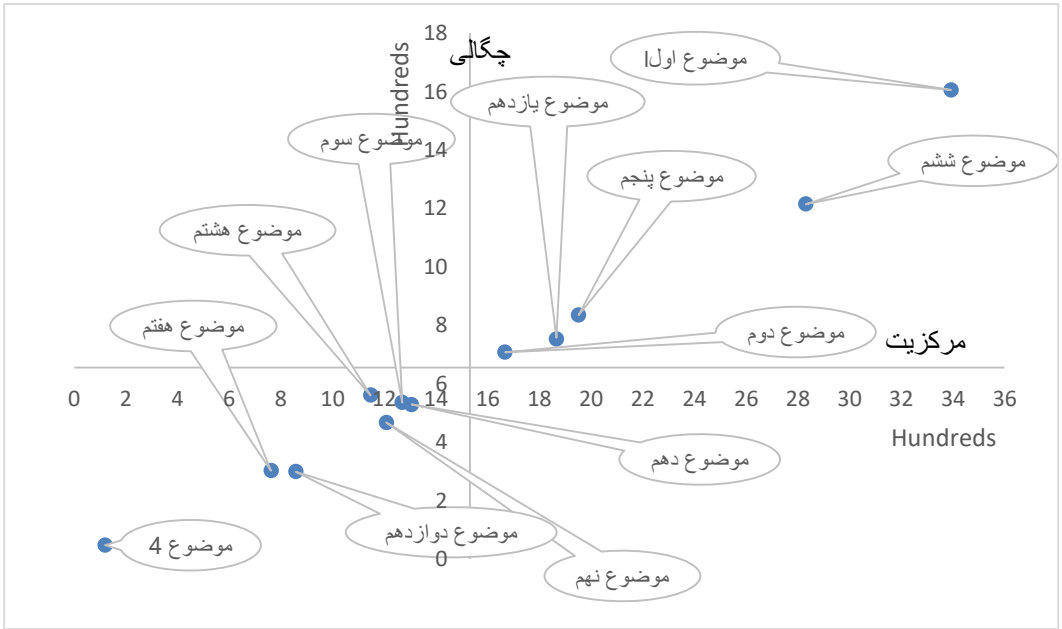


شکل ٤-۲۸ روابط بین موضوع‌ها که منجر به وقوع موضوع ششم می‌شود با استفاده از قوانین انجمنی

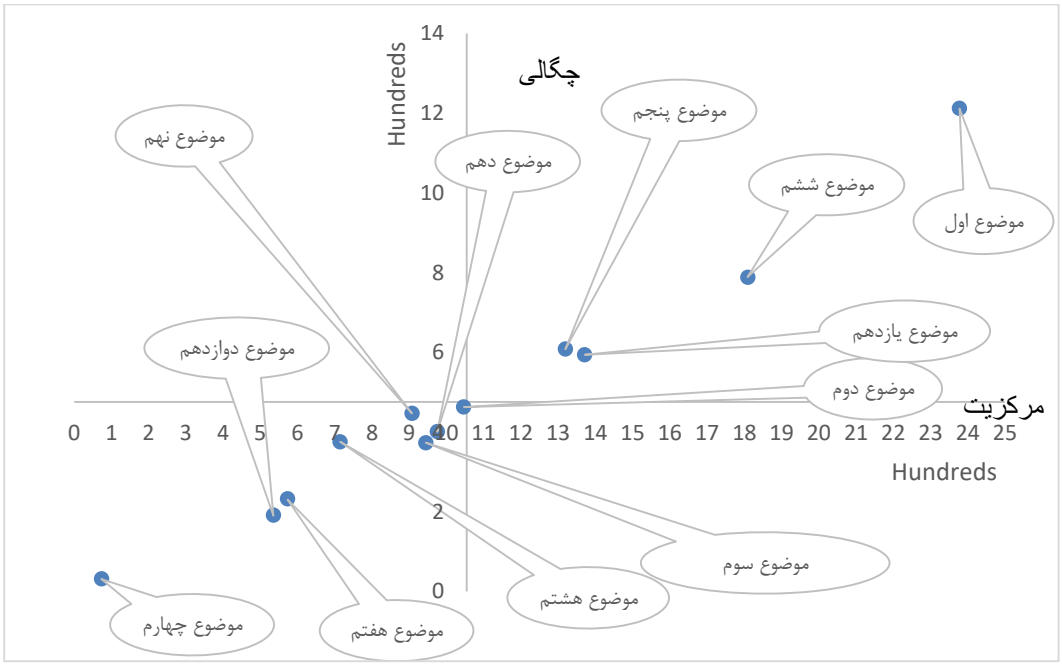
برای تحلیل بیشتر اسناد را در دو بازه زمانی ۱۳۸۸ تا ۱۳۹۱ (بازه زمانی اول) و ۱۳۹۲ تا ۱۳۹۴ (بازه زمانی دوم) تقسیم می‌کنیم و هر کدام را به صورت جداگانه بررسی و میزان مرکزیت و چگالی

هرکدام محاسبه شده‌اند. با مقایسه شاخص‌های این دو بازه زمانی می‌توان روند حرکت موضوع‌ها را مورد بررسی قرار می‌گیرد.

چهارسال اول		چهارسال دوم	
TD	TC	TD	TC
۱۶۰۵	۳۳۹۳	۱۲۱۲	۲۳۷۸
۷۰۸	۱۶۶۶	۴۶۳	۱۰۴۶
۴۶۷	۱۲۰۸	۳۷۳	۹۴۴
۴۸	۱۱۹	۳۱	۷۳
۸۳۵	۱۹۵۱	۶۰۸	۱۳۱۹
۱۲۱۵	۲۹۳۰	۷۸۸	۱۸۰۹
۳۰۳	۷۶۲	۲۳۲	۵۷۳
۵۶۱	۱۱۴۷	۳۷۵	۷۱۴
۵۳۶	۱۲۶۹	۴۴۷	۹۰۷
۵۲۹	۱۳۰۵	۴۰۱	۹۷۵
۷۵۳	۱۸۶۶	۵۹۴	۱۳۷۱
۳۰۰	۸۵۸	۱۹۱	۵۳۵
میانگین			
۱۵۳۱	۶۵۵	۱۰۵۴	۴۶۱



شکل ٤-۲۹ نمودار راهبردی چهار سال اول



شکل ٤-٣٠ نمودار استراتژیک چهارسال دوم

۴-۵-۱-۲- تحلیل موضوع و نمودار راهبردی با استفاده از ماتریس هم‌رخدادی کلمات

در این بخش شاخص‌های مطرح‌شده در فصل سوم را با استفاده از ماتریس هم‌رخدادی به دست آورده و موردبررسی قرار می‌دهیم. این شاخص‌ها میزان مرکزیت موضوع (TC) و میزان چگالی یا بلوغ یک موضوع (TD) را بیان می‌کنند. این شاخص‌ها را با استفاده از روش ارائه‌شده برای نمایش کلمات و با استفاده از شاخص شمول محاسبه‌شده و در نمودارهای مختلف نشان داده خواهند شد. جدول ۴-۷۳ روند میزان هم‌رخدادی کلمات استفاده‌شده در هر خوشه در سال‌های مختلف (هر درایه نماینده شاخص TD و برحسب هم‌رخدادی کلمات در آن سال است) را نشان می‌دهد.

جدول ٤-٧٣ روند میزان هم‌رخدادی کلمات استفاده‌شده در هر خوشه در سال‌های مختلف (هر درایه نماینده شاخص TD و برحسب هم‌رخدادی کلمات در آن سال است)

نام خوشه	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴	۱۳۹۵	تعداد کلمات	مجموع TD یک موضوع
موضوع ۱	۱۵۴۲	۲۴۹۶	۲۶۳۷	۲۶۸۳	۲۸۲۶	۲۴۶۶	۶۲۳	۱۱۵۹	۱۹	۱۶۴۵۱

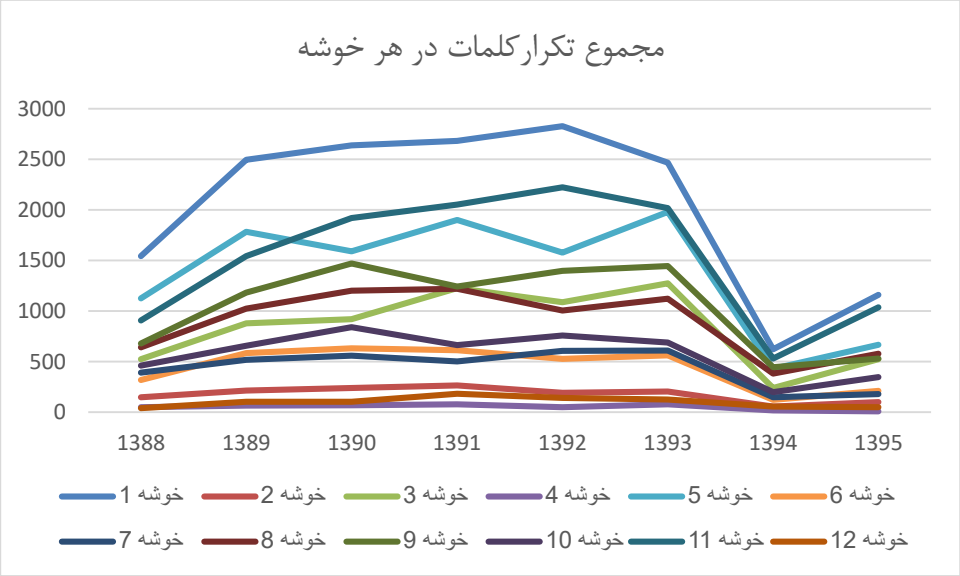
موضوع ۲	۱۴۷	۲۱۵	۲۴۰	۲۶۵	۱۹۱	۲۰۵	۵۶	۱۰۰	۱۴	۱۴۳۳
موضوع ۳	۵۲۵	۸۷۹	۹۲۱	۱۲۲۶	۱۰۸۷	۱۲۷۳	۲۳۹	۵۲۶	۱۰	۶۶۸۶
موضوع ۴	۴۹	۶۵	۶۸	۸۰	۴۹	۷۹	۱۸	۸	۱	۴۱۷
موضوع ۵	۱۱۲۷	۱۷۸۲	۱۵۸۹	۱۹۰۱	۱۵۷۷	۱۹۸۰	۴۲۹	۶۶۶	۱۴	۱۱۰۶۵
موضوع ۶	۳۲۰	۵۸۶	۶۳۳	۶۱۲	۵۲۸	۵۶۳	۱۲۳	۲۱۲	۲۴	۳۶۰۱
موضوع ۷	۳۹۲	۵۱۷	۵۵۸	۵۰۱	۶۰۷	۶۱۰	۱۴۸	۱۸۱	۶	۳۵۲۰
موضوع ۸	۶۴۰	۱۰۲۵	۱۲۰۱	۱۲۲۰	۱۰۰۵	۱۱۲۳	۳۸۲	۵۷۸	۱۱	۷۱۸۵
موضوع ۹	۶۸۰	۱۱۸۱	۱۴۷۱	۱۲۴۳	۱۳۹۷	۱۴۴۴	۴۴۵	۵۳۱	۱۱	۸۴۰۳
موضوع ۱۰	۴۶۰	۶۵۷	۸۴۰	۶۶۳	۷۵۸	۶۸۹	۲۰۰	۳۴۷	۷	۴۶۲۱
موضوع ۱۱	۹۰۷	۱۵۴۲	۱۹۱۸	۲۰۵۴	۲۲۲۳	۲۰۱۷	۵۳۱	۱۰۳۷	۱۸	۱۲۲۴۷
موضوع ۱۲	۴۰	۱۰۵	۱۰۵	۱۸۴	۱۴۳	۱۲۶	۵۶	۴۹	۷	۸۱۵

ازآنجاکه در موضوع‌های مختلف کلمات با فرکانس‌های مختلف استفاده‌شده‌اند و هرچقدر تعداد کلمات خوشه بیشتر باشد تکرار و چگالی کل آن موضوع نیز بیشتر می‌شود، میزان مجموع چگالی کل یک موضوع را به تعداد کلمات موجود در آن موضوع تقسیم کرده و آن را نرمال می‌کنیم. جدول ۴-۷۴روند میزان نرمال شده هم‌رخدادی کل کلمات استفاده‌شده در هر خوشه در سال‌های مختلف (نرمال شده برحسب تعداد کلمات هر خوشه) را نشان می‌دهد.

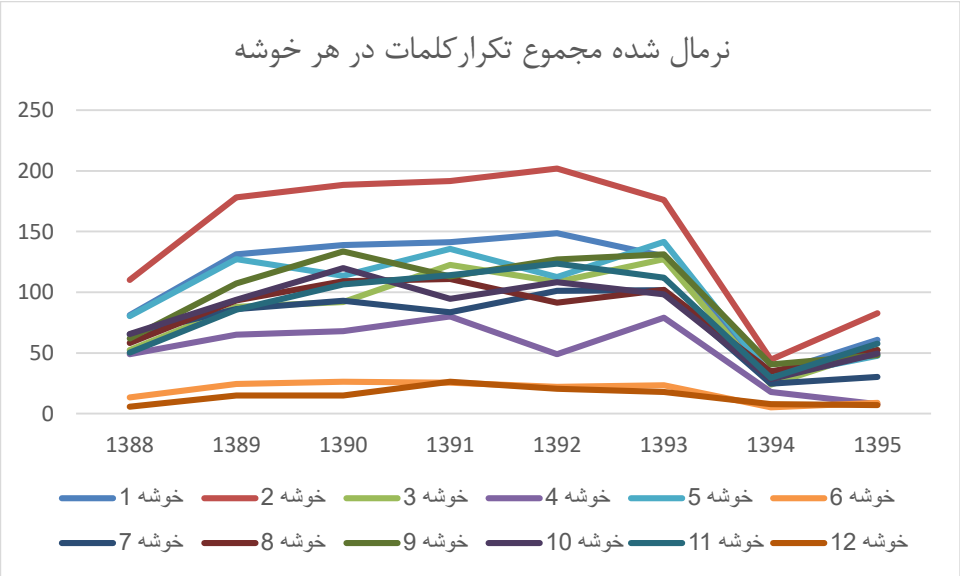
جدول ٤-٧٤روند میزان نرمال شده هم‌رخدادی کل کلمات استفاده‌شده در هر خوشه در سال‌های مختلف (نرمال شده برحسب تعداد کلمات هر خوشه)

	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴	۱۳۹۵
موضوع ۱	۸۱	۱۳۱	۱۳۹	۱۴۱	۱۴۹	۱۳۰	۳۳	۶۱
موضوع ۲	۱۱۰	۱۷۸	۱۸۸	۱۹۲	۲۰۲	۱۷۶	۴۵	۸۳
موضوع ۳	۵۳	۸۸	۹۲	۱۲۳	۱۰۹	۱۲۷	۲۴	۵۳
موضوع ٤	۴۹	۶۵	۶۸	۸۰	۴۹	۷۹	۱۸	۸
موضوع ٥	۸۱	۱۲۷	۱۱۴	۱۳۶	۱۱۳	۱۴۱	۳۱	۴۸
موضوع ٦	۱۳	۲۴	۲۶	۲۶	۲۲	۲۳	۵	۹
موضوع ۷	۶۵	۸۶	۹۳	۸۴	۱۰۱	۱۰۲	۲۵	۳۰
موضوع ۸	۵۸	۹۳	۱۰۹	۱۱۱	۹۱	۱۰۲	۳۵	۵۳

موضوع ۹	۶۲	۱۰۷	۱۳۴	۱۱۳	۱۲۷	۱۳۱	۴۰	۴۸
موضوع ۱۰	۶۶	۹۴	۱۲۰	۹۵	۱۰۸	۹۸	۲۹	۵۰
موضوع ۱۱	۵۰	۸۶	۱۰۷	۱۱۴	۱۲۴	۱۱۲	۳۰	۵۸
موضوع ۱۲	۶	۱۵	۱۵	۲۶	۲۰	۱۸	۸	۷



شکل ٤-۳۱ مجموع تکرار کلمات در هر خوشه در سال‌های مختلف



شکل ٤-۳۲مجموع نرمال شده تکرار کلمات در هر خوشه در سال‌های مختلف

روش دیگر محاسبه میزان روند خوشه‌ها استفاده از مجموع تکرار اسناد استفاده‌شده به ازای هر کلمه است. این مقدار برابر مقدار (DF) هر کلمه است. جدول ۴-۷۵ روند میزان کل اسناد استفاده‌شده به ازای هر کلمه (DF) در هر خوشه در سال‌های مختلف (هر درایه نماینده شاخص TD برحسب هم‌رخدادی کلمات در آن سال است) را نشان می‌دهد.

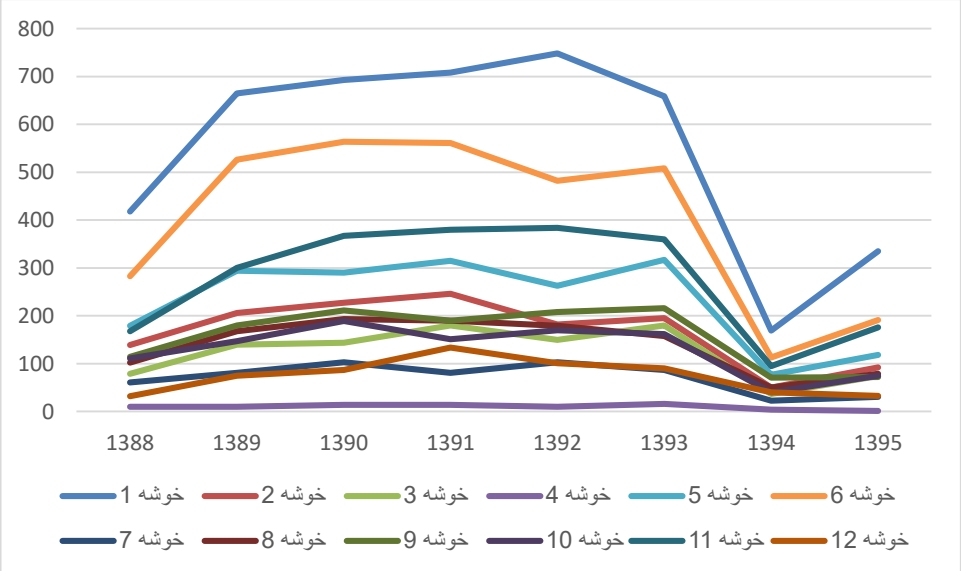
جدول ٤-٧٥ روند میزان کل اسناد استفاده‌شده به ازای هر کلمه (DF) در هر خوشه در سال‌های مختلف (هر درایه نماینده شاخص TD برحسب هم‌رخدادی کلمات در آن سال است)

	۱۳۹۵	۱۳۹۴	۱۳۹۳	۱۳۹۲	۱۳۹۱	۱۳۹۰	۱۳۸۹	۱۳۸۸	
موضوع ۱	۳۳۵	۱۶۹	۶۵۹	۷۴۸	۷۰۸	۶۹۳	۶۶۵	۴۱۸	۱۹
موضوع ۲	۹۲	۵۰	۱۹۵	۱۸۲	۲۴۶	۲۲۷	۲۰۶	۱۳۹	۱۴
موضوع ۳	۷۴	۳۷	۱۸۰	۱۵۰	۱۸۰	۱۴۴	۱۴۰	۷۹	۱۰
موضوع ۴	۱	۴	۱۶	۱۰	۱۴	۱۴	۱۰	۱۰	۱
موضوع ۵	۱۱۸	۷۷	۳۱۷	۲۶۳	۳۱۵	۲۹۰	۲۹۴	۱۷۹	۱۴
موضوع ۶	۱۹۱	۱۱۳	۵۰۸	۴۸۲	۵۶۱	۵۶۴	۵۲۶	۲۸۳	۲۴
موضوع ۷	۳۱	۲۳	۸۷	۱۰۳	۸۱	۱۰۳	۸۱	۶۱	۶
موضوع ۸	۷۹	۵۰	۱۵۸	۱۷۹	۱۹۰	۱۹۳	۱۶۸	۱۰۲	۱۱
موضوع ۹	۷۲	۷۱	۲۱۶	۲۰۸	۱۹۰	۲۱۱	۱۸۰	۱۱۴	۱۱
موضوع ۱۰	۷۵	۴۰	۱۶۲	۱۷۰	۱۵۱	۱۸۹	۱۴۷	۱۱۱	۷
موضوع ۱۱	۱۷۶	۹۵	۳۶۰	۳۸۴	۳۸۰	۳۶۷	۳۰۰	۱۶۸	۱۸
موضوع ۱۲	۳۳	۴۰	۹۰	۱۰۱	۱۳۴	۸۷	۷۵	۳۲	۷

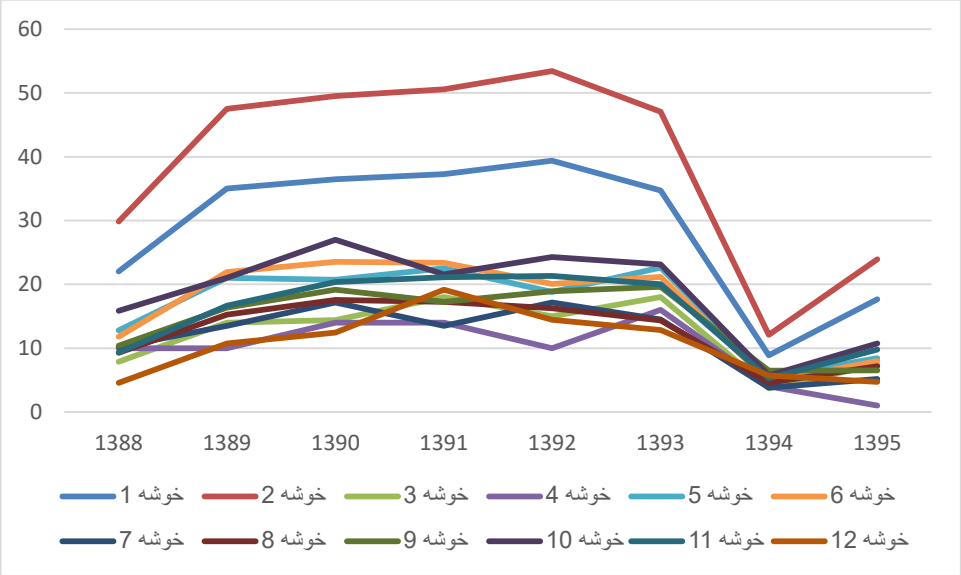
ازآنجاکه در موضوع‌های مختلف کلمات با فرکانس‌های مختلف استفاده‌شده‌اند و هرچقدر تعداد کلمات خوشه بیشتر باشد تکرار و چگالی کل آن موضوع نیز بیشتر می‌شود، میزان مجموع چگالی کل یک موضوع را به تعداد کلمات موجود در آن موضوع تقسیم کرده و آن را نرمال می‌کنیم. جدول ۴-۷۶ روند نرمال شده میزان کل اسناد استفاده‌شده به ازای هر کلمه (DF) در هر خوشه در سال‌های مختلف را نشان می‌دهد.

جدول ٤-٧٦ روند نرمال شده میزان کل اسناد استفاده‌شده به ازای هر کلمه (DF) در هر خوشه در سال‌های مختلف

	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴	۱۳۹۵
موضوع ۱	۲۲	۳۵	۳۶	۳۷	۳۹	۳۵	۹	۱۸
موضوع ۲	۳۰	۴۸	۵۰	۵۱	۵۳	۴۷	۱۲	۲۴
موضوع ۳	۸	۱۴	۱۴	۱۸	۱۵	۱۸	۴	۷
موضوع ۴	۱۰	۱۰	۱۴	۱۴	۱۰	۱۶	۴	۱
موضوع ۵	۱۳	۲۱	۲۱	۲۳	۱۹	۲۳	۶	۸
موضوع ۶	۱۲	۲۲	۲۴	۲۳	۲۰	۲۱	۵	۸
موضوع ۷	۱۰	۱۴	۱۷	۱۴	۱۷	۱۵	۴	۵
موضوع ۸	۹	۱۵	۱۸	۱۷	۱۶	۱۴	۵	۷
موضوع ۹	۱۰	۱۶	۱۹	۱۷	۱۹	۲۰	۶	۷
موضوع ۱۰	۱۶	۲۱	۲۷	۲۲	۲۴	۲۳	۶	۱۱
موضوع ۱۱	۹	۱۷	۲۰	۲۱	۲۱	۲۰	۵	۱۰
موضوع ۱۲	۵	۱۱	۱۲	۱۹	۱۴	۱۳	۶	۵



شکل ٤-٣٣نرمال شده مجموع اسناد استفاده‌شده در هر خوشه



شکل ٤-٣٤نرمال شده مجموع اسناد استفاده‌شده در هر خوشه

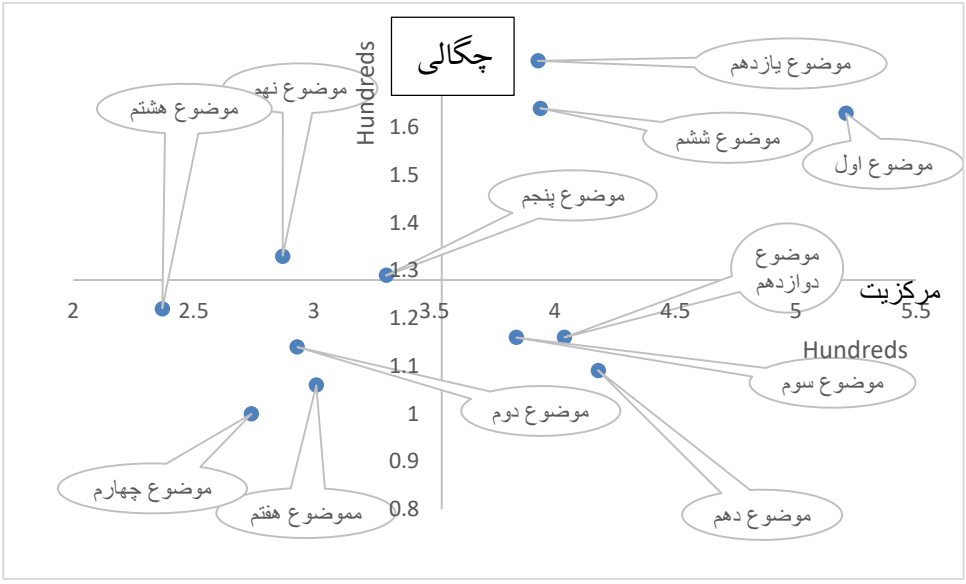
جدول ۴-۷۷ روند میزان اسناد استفاده‌شده در یک موضوع در سال‌های مختلف را نشان می‌دهد.

روند رشد موضوع‌ها بر اساس تعداد اسناد چاپ‌شده در آن موضوع نشان داده شده است.

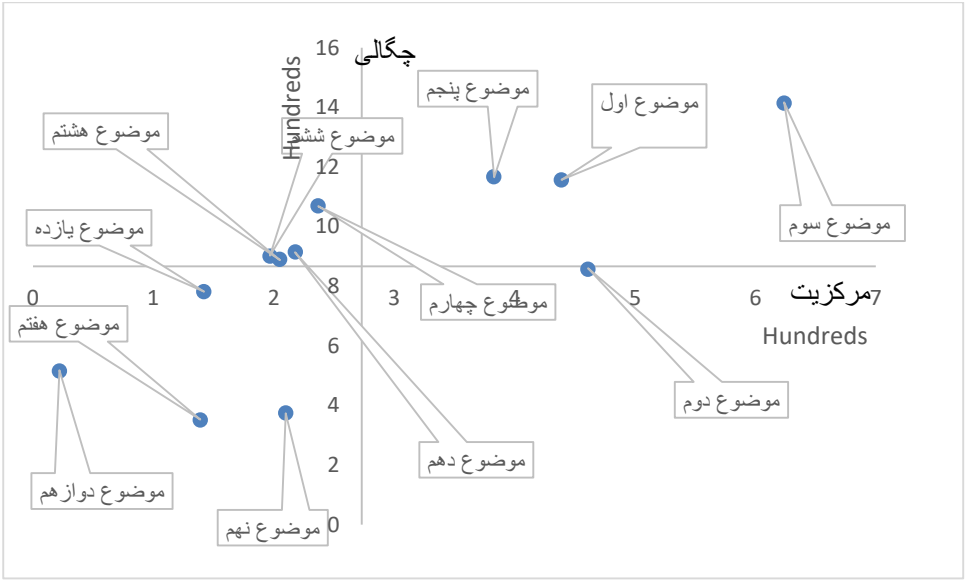
جدول ٤-۷۷ روند میزان اسناد استفاده‌شده در یک موضوع در سال‌های مختلف

	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹٤	۱۳۹٥
موضوع ۱	۲۷۷	۴۱۲	۴۶۶	۴۵۰	۴۶۸	۴۱۹	۱۱۱	۲۱۴
موضوع ۲	۱۲۱	۱۸۱	۱۹۲	۲۱۴	۱۶۹	۱۷۴	۴۲	۷۸
موضوع ۳	۷۴	۱۱۳	۱۲۹	۱۵۱	۱۲۵	۱۵۱	۳۱	۶۶
موضوع ٤	۱۰	۱۰	۱۴	۱۴	۱۰	۱۶	۴	۱
موضوع ٥	۱۴۱	۲۲۱	۲۲۴	۲۴۹	۲۰۵	۲۵۰	۶۲	۹۱
موضوع ٦	۲۰۱	۳۱۶	۳۴۷	۳۵۱	۲۹۱	۳۲۰	۵۵	۱۲۲
موضوع ۷	۵۳	۷۴	۹۹	۷۷	۹۶	۸۴	۲۲	۳۰
موضوع ۸	۸۹	۱۴۶	۱۷۰	۱۵۶	۱۴۲	۱۳۴	۳۸	۶۱
موضوع ۹	۹۰	۱۲۹	۱۶۳	۱۵۴	۱۶۷	۱۷۰	۵۵	۵۵
موضوع ۱۰	۹۶	۱۲۷	۱۷۰	۱۳۶	۱۵۴	۱۴۸	۳۱	۶۸
موضوع ۱۱	۱۱۸	۱۹۸	۲۲۰	۲۱۷	۲۲۳	۲۱۷	۴۸	۱۰۶
موضوع ۱۲	۴۲	۷۳	۹۴	۹۱	۸۶	۷۱	۷	۲۷

شکل ۴-۳۵ نمودار راهبردی برحسب شاخص شمول و شکل ۴-۳۶ نمودار راهبردی برحسب شاخص ارائه‌شده در این رساله را نشان می‌دهد.



شکل ٤-۳۵ نمودار راهبردی برحسب شاخص شمول



شکل ٤-۳٦ نمودار راهبردی برحسب شاخص ارائه‌شده

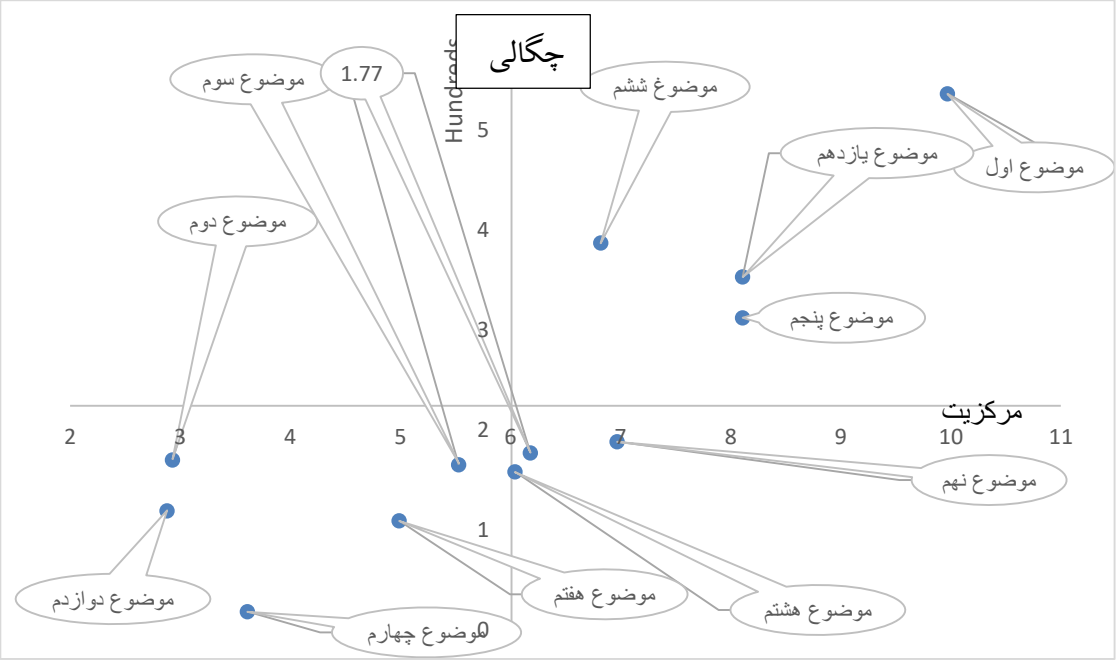
جدول ٤-٧٨ شاخص‌های مرکزیت و چگالی برحسب میزان شاخص‌های شمول و ارائه‌شده در این رساله

	TC_Inclusion	TD_Inclusion	TC Porposed	TD Proposed
موضوع ۱	۵۲۱	۱۶۳	۶۲۴	۱۴۱۴
موضوع ۲	۲۹۳	۱۱۴	۲۱۰	۳۷۴
موضوع ۳	۳۸۴	۱۱۶	۲۰۵	۸۹۰
موضوع ٤	۲۷۴	۱۰۰	۲۲	۵۱۵
موضوع ٥	۳۳۰	۱۲۹	۳۸۳	۱۱۶۷
موضوع ٦	۳۹۴	۱۶۴	۴۶۱	۸۵۷
موضوع ۷	۳۰۱	۱۰۶	۱۴۲	۷۸۱
موضوع ۸	۲۳۷	۱۲۲	۱۹۷	۹۰۱
موضوع ۹	۲۸۷	۱۳۳	۲۳۷	۱۰۶۹
موضوع ۱۰	۴۱۸	۱۰۹	۲۱۸	۹۱۵
موضوع ۱۱	۳۹۳	۱۷۴	۴۳۹	۱۱۵۶
موضوع ۱۲	۴۰۴	۱۱۶	۱۳۹	۳۵۱
میانگین	۳۵۳	۱۲۹	۲۷۳	۳۵۱

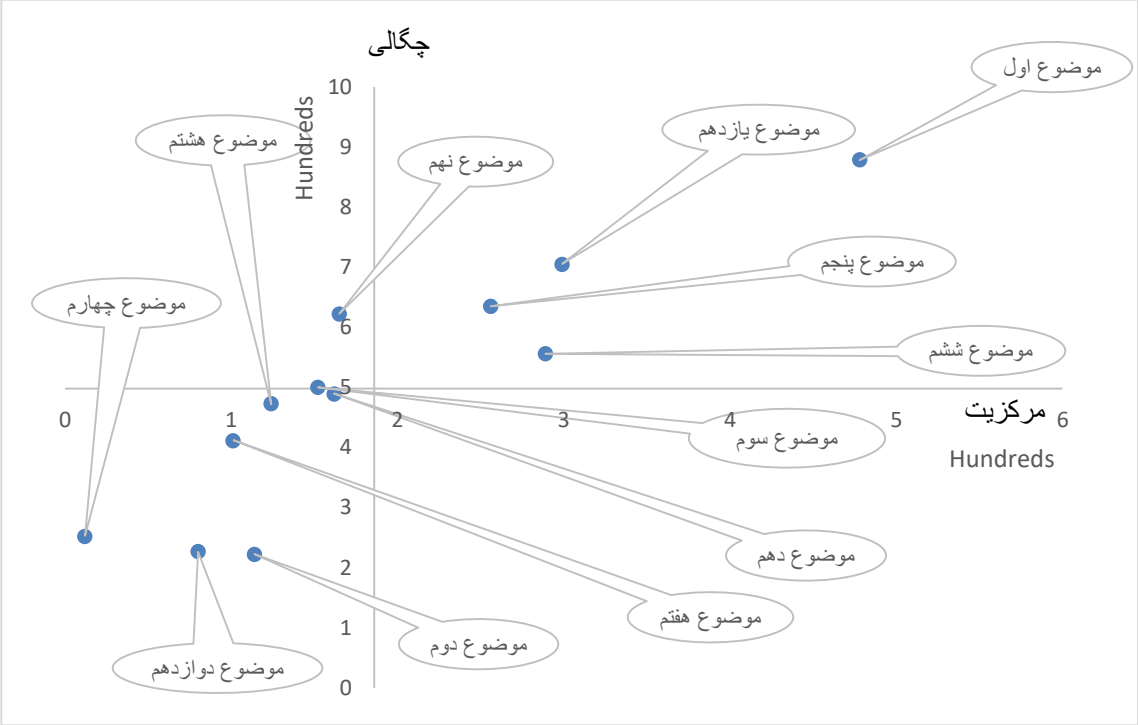
جدول ٤-٧٩ مرکزیت و بلوغ موضوع‌ها در دو بازه زمانی چهار ساله بر اساس ماتریس هم‌رخدادی موضوع‌ها

چهارسال اول		چهارسال دوم	
TD	TC	TD	TC
۵۳۶	۹۹۷	۸۷۹	۴۷۸
۱۷۰	۲۹۳	۲۲۲	۱۱۴
۱۶۵	۵۵۳	۵۰۰	۱۵۲
۱۸	۳۶۱	۲۵۲	۱۲
۳۱۲	۸۱۱	۶۳۵	۲۵۶
۳۸۷	۶۸۲	۵۵۶	۲۸۹
۱۰۹	۴۹۹	۴۱۱	۱۰۱
۱۵۸	۶۰۴	۴۷۳	۱۲۴
۱۸۸	۶۹۷	۶۲۲	۱۶۵

۱۶۲	۴۸۹	۶۱۸	۱۷۷
۲۹۹	۷۰۵	۸۱۱	۳۵۳
۸۰	۲۲۷	۲۸۸	۱۱۷
میانگین			
۴۹۸	۱۸۶	۲۲۴	۶۰۱



شکل ٤-۳۷نمودار راهبردی چهار سال اول بر اساس کلیدواژه‌ها



شکل ٤-٣٨نمودار راهبردی چهار سال دوم بر اساس کلیدواژه‌ها

۴-۵-۲- تحلیل درون موضوع

ازآنجاکه داخل هر موضوع می‌تواند به‌صورت جداگانه تحلیل شود و کلمات اجزای تشکیل‌دهنده هر موضوع هستند، اطلاعات آماری می‌بایست از کلمات استخراج شوند. جدول ۴-۸۰ تحلیل درون موضوع‌ها بر مبنای فراوانی و نرخ رشد و نزول هر موضوع بر اساس تکرار کلمات را نشان می‌دهد. لازم به ذکر است که در این جدول مجموعه‌کلمات، یک موضوع را نشان می‌دهند و توسط شخص متخصص می‌بایست بر اساس همه کلمات موضوع را نام گذاری کند. از آنجا که بخاطر خطای نرمالساز و برچسب زدن کلمات وجود دارد، برخی کلمات ممکن است از فیلترهای پیشنهادی عبور کنند. برای این منظور همانطور که شکل ۱-۱ نشان داده شده است برای برطرف کردن مشکلات بوجود آمده شخص متخصص می‌تواند ویرایش‌هایی در موضوع‌ها و کلیدواژه‌های هر موضوع انجام دهد.

در هر خوشه پارامترهای زیر محاسبه خواهند شد:

- کلماتی که بیشترین میزان تأثیر **(بیشترین فراوانی هم‌رخدادی)** را در فراوانی خوشه دارند.

- کلماتی که **بیشترین رشد** را در خوشه دارند
- کلماتی که بیشترین **نرخ کاهش فراوانی** رادارند.

جدول ٤-٨٠ تحلیل درون موضوع‌ها بر مبنای فراوانی و نرخ رشد و نزول هر موضوع بر اساس تکرار کلمات

	کلمات با بیشترین فراوانی	کلمات با بیشترین میانگین نرخ رشد در طول زمان	کلمات با بیشترین میانگین نرخ کاهش در طول زمان
موضوع ۱	الگوریتم، شبکه، بهبود، بررسی، تحلیل، بهینه‌سازی، عملکرد، جریان	الگوریتم، بهینه‌سازی، شبکه، بررسی	مدل‌سازی، پیاده‌سازی، تحلیل، جریان، بهبود
موضوع ۲	پهنای‌نوار، استخراج ویژگی، تبدیل موجک، پردازش تصویر، حسگربی‌سیم، پردازش سیگنال	پردازش تصویر، پردازش سیگنال، سیستم‌غیرخطی	ماشین‌بردار پشتیبان، استخراج‌ویژگی، کاربرد، فناوری‌سی‌ماس
موضوع ۳	ریزشبکه، ژنراتور، بازار، نیروگاه، ترانسفورماتور، پایداری	نیروگاه	ریزشبکه، حفاظت، مبدل
موضوع ٤	تبدیل	تبدیل	-
موضوع ٥	تشخیص، خطا، فازی، مدل، تخمین، شناسایی، سیگنال	تطبیقی، فازی، سیگنال	تشخیص، فیلتر، خطا، مدل
موضوع ٦	شبکه عصبی، سیستم قدرت، توربین بادی، تولید پراکنده، قابلیت اطمینان، الگوریتم ژنتیک، شبکه توزیع	شبیه‌سازی رایانه‌ای، توربین‌بادی، تولید پراکنده، شبکه عصبی، سیستم قدرت	قابلیت اطمینان، شبکه‌هوشمند، توزیع نیروی برق
موضوع ۷	کانال، زمان، فرکانسی، سرعت، سنکرون	کانال، فرکانسی	-
موضوع ۸	آنتن، تصویر، نوری، آشکارسازی، ردیابی	ردیابی، آنالوگ، آشکارسازی	تصاویر، آنتن، سیگنال

موضوع ۹	باند، حسگر، موتور، نویز	موتور	فاز، نویز
موضوع ۱۰	فرکانس، ساختار، نظر	فرکانس، مدل‌سازی	ساختار، ارائه
موضوع ۱۱	تولید، انرژی، منابع، برق، تلفات	ظرفیت، خطوط، مدیریت، انتقال، منابع	انرژی، اطمینان، تلفات
موضوع ۱۲	تأثیر، بازار برق، عدم قطعیت، شبیه‌سازی سیستم‌قدرت	تأثیر، عدم قطعیت، شبیه‌سازی سیستم‌قدرت	–

در جدول ۴-۸۰ تحلیل درون موضوع‌ها بر مبنای فراوانی و نرخ رشد و نزول هر موضوع بر اساس تکرار کلمات انجام‌شده است. با استفاده از میزان تکرار کلمات در هر موضوع در سال‌های متوالی می‌توان روند رشد کلمات و نزول آن‌ها را محاسبه کرد. ستون کلمات با بیشترین تکرار، کلماتی را نشان می‌دهد که در بازه زمانی ۱۳۸۸ تا ابتدای ۱۳۹۶ در پارساها ثبت‌شده در پایگاه گنج وجود داشته است. دو ستون دیگر نیز با استفاده از محاسبه میانگین تغییرات تکرار کلیدواژه‌ها در بازه زمانی ذکرشده در سال‌های مختلف، میزان رشد و نزول یک کلمه محاسبه‌شده و برترین کلیدواژه‌ها در درایه مربوطه در جدول آمده است.

فصل ۵- تحلیل، جمع‌بندی و کارهای آینده

۵-۱- مقدمه

انجام هر نوع فعالیت علمی و عملی در یک حوزه خاص نیازمند آن است تا درک و فهم مناسبی از پژوهش‌های انجام شده در آن حوزه و روند تحقیقاتی آن حاصل شود. بنابراین بررسی روند تحقیقات یک حوزه علمی (در بازه‌های زمانی مختلف) می‌تواند درک بهتری را برای محققین و سیاست‌گذاران آن حوزه ایجاد نماید تا بتوانند برنامه‌ریزی مناسبی را جهت انجام تحقیقات آتی و تخصیص منابع پژوهشی داشته باشند. یکی از مهمترین رویکردها در تحلیل‌روند یک حوزه علمی، بررسی اسناد علمی منتشرشده در آن حوزه با استفاده از روش‌های علم‌سنجی و پیمایش اطلاعات و متون اسناد است. بنابراین تحلیل‌روند ابزار مناسب برای محققین و سیاست‌گذاران در اجرای فعالیت‌های آن‌ها است. از اینرو انتخاب روشی مناسب برای تحلیل وضعیت فعلی و پیش‌بینی آن حائز اهمیت است.

با توجه به این نکته سامانه و روشی برای بررسی و تحلیل‌روند پارسا فارسی که به صورت ماشینی عمل کند در مطالعات یافت نشده است، این خلع مارا برآن داشت تا در این رساله با بررسی جنبه‌های مختلف روش‌های تحلیل‌روند، رویکردی جامع برای تحلیل‌روند پایان‌نامه‌ها و رساله‌های فارسی ارائه دهیم. از آنجا که دقت و جامعیت تحلیل‌روند از اهمیت ویژه‌ای برخوردار است، رویکرد ارائه‌شده در این رساله از روش‌های متن‌کاوی و اطلاعات کتابشناختی پارسا استفاده کرده است، تا بتوان علاوه بر تفکیک و دسته‌بندی مناسب حوزه‌ها، روند پژوهشی در هرکدام از حوزه‌ها نیز استخراج شود.

بررسی پیشینه نشان داد که روش‌های مختلفی برای کشف حوزه‌های علمی و تحلیل‌روند آن‌ها وجود دارد. عموم روش‌های کشف حوزه‌ها و تحلیل‌روند شامل چند گام مشخص و یکسان هستند.

این مراحل عبارتند از: استخراج اسناد، استخراج کلیدواژه‌ها، تشکیل ماتریس هم‌رخدادی، خوشه‌بندی کلمات، موضوع‌بندی و تحلیل موضوع‌ها. هر کدام از گام‌های بیان شده خود شامل روش‌ها و عملکردهای مختلفی هستند که خود می‌توانند بر عملکرد روش تحلیل‌روند تاثیر بگذارند. بنابراین انتخاب، استفاده یا ارائه راهکاری که بهترین عملکرد را در حوزه پارسای فارسی داشته‌باشد در این رساله اهمیت ویژه‌ای دارد. در این رساله با مطالعه و بررسی روش‌های پیشین در هر گام، بهترین روش انتخاب شده یا روشی ارائه شده که عملکرد بهتری از روش‌های پیشین داشته باشد.

۵-۲- خلاصه دانش‌افزایی

انواع نوآوری را به می توان سه دسته تقسیم کرد: نوآوری در روش، نوآوری در موضوع و موضوع در بافت (زمینه‌پژوهش). در این رساله برای نخستین بار طرحی برای تحلیل‌روند با استفاده از یادگیری‌ماشین و به صورت تمام خودکار برای متون علمی فارسی ارائه‌شده است. علاوه بر آن در گام‌های تحلیل‌روند نوآوری‌هایی در روش انجام ارائه شده تا هر گام از تحلیل‌روند به صورت بهینه عمل کند. در انتها نیز روشی برای ترسیم ترسیم نمودار راهبردی بمنظور تحلیل‌روند با دیدی کلانتر به موضوع‌ها ارائه‌شده است. روش پیشنهادی از بخش‌های مختلفی تشکیل شده است. هر بخش به‌صورت جداگانه موردبررسی قرار گرفت و نوآوری هر بخش به‌تفصیل در همان بخش مربوطه بیان گردید.

در گام نخست به شناخت خصوصیات اسناد در دسترس پرداخته شده است. روابط بین انواع کلیدواژه‌ها استخراج شده و توزیع آن‌ها به منظور استخراج خودکار کلیدواژه‌ها در عنوان و چکیده مورد بررسی قرار گرفته شده است. برای این منظور اسناد مورد نیاز استخراج و ساختار چکیده‌ها و کلیدواژه‌های پارسا را مورد بررسی و ارزیابی داده‌ایم. سپس اسناد را برای پردازش‌های زبان طبیعی باید به شکل مناسب و قابل پردازش برای سیستم‌های دیجیتال تبدیل کنیم. بنابراین در این بخش روشی برای استخراج کلیدواژه‌های اسناد ارائه کرده‌ایم. سپس با استفاده از روش پیشنهادی برای بازنمایی کلمات، کلیدواژه‌ها را به شکلی موثر برای استخراج موضوع‌های اسناد آماده ساخته‌ایم. در گام بعدی با استخراج الگوهای رفتاری بین کلمات (با استفاده از الگوریتم‌های کا-میانه و خوشه‌بندی فازی) موضوع‌های مخفی موجود در متون را مشخص خواهیم کرد.

در این رساله چند رویکرد جدید برای تحلیل‌روند با استفاده از نمودار راهبردی و قوانین انجمنی پیشنهاد شده است. علاوه‌بر محاسبه شاخص‌ها با استفاده از روش‌های متداول، شاخص تکمیلی که نقطه ضعف دخیل نکردن میزان تکرار کلمه را برطرف کرده است ارائه شده است. علاوه بر آن در رویکرد پیشنهادی محاسبه شاخص‌ها و ترسیم نمودار راهبردی بر اساس ماتریس هم‌رخدادی موضوعی محاسبه و نمایش داده شده‌اند. روش پیشنهادی هم‌رخدادی موضوعی باعث حذف کلمات نویز (کلماتی موجود در سند علمی که موضوع اصلی را بیان نمی‌کنند) تحلیلی با دید کلان تر (تحلیل از سطح کلمات علمی به سطح اسناد علمی) شده است.

روشی دیگری که برای تحلیل‌روند در این رساله ارائه شده است، استفاده از قوانین انجمی برای استخراج روابط بین موضوع‌ها و پیش‌بینی استفاده از موضوع‌ها است. علاوه‌بر استفاده از نمودار راهبردی برای بصری‌سازی و تحلیل، روابط بین موضوع‌ها را با استفاده از روش پیشنهادی قوانین انجمنی استخراج شده‌اند. با استفاده از استخراج این روابط می‌توان وابستگی چندگانه موضوع‌ها به یکدیگر را ملاحظه کرد. بر اساس این روابط احتمال استفاده (میزان اطمینان بدست آمده از قواعد انجمنی) هر یک از موضوع‌ها توسط یک پژوهشگر مشخص می‌شود و می‌توان پژوهش‌های آتی افراد را پیش‌بینی کرد.

در بخش بعد روش‌های ارائه‌شده پیاده‌سازی و موردبحث و بررسی قرار خواهند گرفت.

۵-۳- چکیده یافته‌های پژوهش و پاسخ به پرسش‌های پژوهش

برای رسیدن به روش تحلیل‌روند با چند پرسش مواجه شدیم که پاسخ به مجموع این پرسش‌ها ما را به روشی برای انجام یک تحلیل مناسب رساند. در ادامه با پاسخ به هر یک از این پرسش‌ها به ارائه نتایج و جمع‌بندی در این خصوصمی‌پردازیم.

۵-۳-۱- پرسش نخست: ساختار چیکده‌های پارساها به چه صورت است؟

بیشتر پژوهش‌های پیشین در زمینه مقایسه کلیدواژه‌ها و توصیفگرها بر روی مقالات انجام‌شده است. در این پژوهش به بررسی میزان تشابه کلیدواژه‌ها و توصیفگرها پرداخته و نشان داده‌شده است که شباهت بین کلیدواژه‌های نویسنده و توصیفگرهای نمایه‌ساز کمتر از ۸٪ است. اختلاف این دو

نشان دهنده اختلاف نظر زیاد در انتخاب کلیدواژه برای متن بین نویسنده و نمایه‌ساز است. بنابراین نمی‌توان به تنهایی به یکی از این نمایه‌ها در تحلیل‌ها اتکا کرد. این نتیجه مطابق با نتایج پژوهش «قنواتی» در سال ۲۰۱۸ است که برروی مقاله‌های پایگاه‌اطلاعاتی اریک و مندلی انجام شد و نشان داد که شباهت بین توصیفگرها و کلیدواژه‌ها ۴٪ است .

جدول ۴-۳ و جدول ۴-۴ نشان می‌دهد که در ۶۰٪ رساله‌ها کمتر از ۲۵٪ کلیدواژه‌های آن‌ها از عنوان انتخاب شده‌اند. جدول ۴-۶ و جدول ۴-۵ نشان می‌دهد که تنها ۱۰٪ از نمایه‌ها از داخل چکیده و عنوان انتخاب نشده‌اند و حدود ۵۰٪ از اسناد کلیدواژه‌های خود را عینا از داخل چکیده استخراج کرده‌اند که در پژوهش (Strader۲۰۰۹) نیز نتیجه مشابه ۵۴٪ گرفته شده است.

در ادامه میزان اشتراک نمایه‌های استفاده‌شده توسط نویسنده و نمایه‌ساز با عنوان و چکیده موردبررسی قرار گرفت. بررسی انجام‌شده نشان داد که کلیدواژه‌های نویسنده در ۶۲٪ از پارساها کمتر از ۲۵٪ با عنوان استفاده کرده‌اند که این مقدار در مقایسه با انتخاب ۷۰٪ کلیدواژه‌ها از عنوان در پژوهش «انصاری» بسیار کمتر است (Ansari۲۰۰۵). از طرفی دیگر ۵۱٪ از کلیدواژه‌های پارساهای بررسی‌شده بیش از ۷۵٪ با چکیده و عنوان شباهت دارند که در پژوهش (Strader۲۰۰۹) نیز نتیجه ۵۴٪ی گرفته شده است این نتایج نشان می‌دهند که نویسندگان عموماً اکثر کلیدواژه‌های خود را عیناً از چکیده انتخاب می‌کنند تا عنوان. از آنجا که عنوان یک رساله فهم مهمی از یک سند علمی را نشان می‌دهند و توصیه شده است که برای تحلیل‌های اسناد از عنوان استفاده شود (He۱۹۹۹) و (.Lv et al۲۰۱۱) لذا در این رساله علاوه بر استفاده از کلیدواژه‌های انتخاب شده توسط نمایه‌سازی و نویسنده از کلیدواژه‌های عنوان نیز استفاده شده است.

جدول ۴-۸ در مقایسه با استاندارد چکیده‌نویسی انسی/نیزو نشان می‌دهد که حدود ۵۵٪ از این پارس‌ها حداکثر اندازه چکیده را رعایت نکرده‌اند و ۳۸٪ از پارس‌ها نیز تعداد کلیدواژه‌های انتخاب کرده را خارج از راهنماهای ناشران و دانشگاه‌ها برای نویسندگان انتخاب نموده‌اند. که می‌توان نتیجه گرفت که تعداد زیادی از دانشجویان تحصیلات تکمیلی از استانداردهای نوشتن چکیده و انتخاب کلیدواژه اطلاع کافی ندارند و یا آن را رعایت نمی‌کنند.

در انتها به بررسی وضعیت توزیع نمایه‌ها در داخل چکیده پرداخته و نشان داده شد ۶۴٪ نمایه‌هایی که در داخل چکیده وجود دارند در ۴۰٪ بخش نخست هستند. این آمار نشان می‌دهد که نویسنده بیشتر مطالب خود را در هدف و روش‌شناسی بیان می‌کند و با استفاده از این ویژگی سیستم‌های استخراج خودکار کلیدواژه و یا خلاصه‌سازهای متن می‌توانند به کلمات کاندیدی که به‌عنوان کلیدواژه از این بخش‌ها استخراج می‌کنند وزن بیشتری قرار دهند. در این بخش نشان داده‌شده است تمام کلیدواژه‌های موجود در چکیده توسط نمایه‌ساز و دانشجو انتخاب شده‌اند اما از آنجا که کلیدواژه‌های موجود در عنوان به صورت کامل مورد انتخاب قرار نگرفته‌اند می‌توان از این ترکیب کلیدواژه‌های عنوان و نمایه‌های موجود در تحلیل‌روند بهره گرفت.

۵-۳-۲- پرسش دوم و سوم: چه روشی برای استخراج و پالایش کلمات‌کلیدی مناسب است؟ و کدام دسته‌بندی برای دسته‌بندی پارسای فارسی مناسب است؟

روش‌های مختلفی برای استخراج و پالایش کلمات‌کلیدی وجود داردر روش فیلترینگ از پارامترهای مختلفی استفاده می‌شود که در بخش ۲-۱-۵- به آن اشاره شده است (فیلترهای استفاده شده در این رساله عبارتند از: TF، DF، POS، حذف کلمات ایست). همانطور که می‌دانیم پالایش ویژگی‌ها (که در اینجا کلیدواژه‌ها هستند) تا زمانی ادامه پیدا می‌کند تا به بازده مناسبی برسیم.

همانطور که در جدول ۴-۹ ملاحظه می‌کنید با استفاده از روش ارائه‌شده در بخش سوم در انتخاب خودکار کلیدواژه‌های کاندید باعث کاهش ۴ تا ۳۷ درصدی کلیدواژه‌های کاندید در هر یک از رشته‌های انتخاب شده است. اما در هنگامی که چند رشته با هم ترکیب می‌شوند تعداد کلیدواژه‌ها بسیار بالاتر می‌روند، با استفاده از روش ارائه‌شده همانطور که در جدول ۴-۱۰ ملاحظه می‌کنید کلیدواژه‌های بسیاری که تعداد تکرار یکسانی دارند از کلیدواژه‌های کاندید حذف شده‌اند. این کاهش بین ۵۴ تا ۷۸ درصد از کلیدواژه‌های کاندید را شامل شده است (در جداول بخش ۴-۳- دسته‌بندی اسناد ملاحظه می‌شود که با وجود کاهش بالای کلیدواژه‌ها دقت دسته‌بندی نزدیک به ۹۹ درصد است که نشان دهنده عملکرد مناسب روش ارائه‌شده در انتخاب کلیدواژه در بخش ۳-۳- دارد). از طرفی دیگر با کاهش ۵۴ تا ۷۸ درصدی کلیدواژه‌ها توسط رابطه ارائه‌شده در این

رساله، می‌توان ملاحظه کرد که میزان صحت دسته‌بندی تنها در حدود ۳ درصد کاهش پیدا کرده است.

برای اینکه آزمودن عملکرد کلیدواژه‌ها، با استفاده از کلیدواژه‌های استخراج شده مجموعه‌داده‌هایی را از رشته های مختلف تشکیل داده‌ایم و با استفاده از انواع دسته‌بندی‌ها آن‌ها را مورد آزمون قرار داده‌ایم. جدول ۴-۱۳، جدول ۴-۱۵ و جدول ۴-۱۷ نتایج دسته‌بندی با استفاده از دسته‌بندهای بیزین، جنگل تصادفی، ترکیبی بگینگ، رگرسیون لجستیک که توسط کتابخانه وکا پیاده‌سازی شده‌اند را نشان می‌دهند. این نتایج نشان می‌دهد که از بین چهار دسته‌بند مورد ارزیابی، دسته‌بند جنگل تصادفی بهترین عملکرد (به طور میانگین ۹۹ ٪ درصد صحت در دسته‌بندی) را در سه آزمون انجام شده داشته است. بعد از این الگوریتم بگینگ با دقت میانگین ۹۴ درصد در رتبه دوم الگوریتم‌های دسته‌بندی قرار گرفته است.

به عنوان سهم علمی رساله در قسمت اول این بخش روشی برای پالایش بیشتر کلیدواژه‌های کاندید به صورت خودکار ارائه‌شده است که با کاهش ۵۰ تا ۷۰ درصدی کلیدواژه‌ها تنها حداکثر تا دو درصد دسته‌بندی کاهش پیدا کرده است. سهم علمی دیگری که در قسمت دوم این بخش ارائه گردید انتخاب بهترین روش دسته‌بندی بر روی اسناد علمی فارسی بوده است که در مطالعات پیشین یافت نشده است. لازم به ذکر است که در بخش بعد با استفاده از روش ارائه‌شده در نمایش کلمات و تبدیل ماتریس سند-کلمه به ماتریس سند-موضوع نوعی استخراج ویژگی صورت پذیرفته شده است که منجر به افزایش دقت در الگوریتم‌های خوشه‌بندی شده است.

۵-۳-۳- پرسش چهارم: چه روشی برای خوشه‌بندی رساله‌ها مناسب است؟

در الگوریتم‌های خوشه‌بندی داده‌های داخل یک خوشه نسبت به داده‌های خوشه‌های دیگر، شباهت بیشتری با یکدیگر دارند. بنابراین بعد از خوشه‌بندی اسناد هرچقدر اسناد داخل یک خوشه به هم مرتبط‌تر(دریک حوزه علمی) باشند، خوشه‌بندی بهتری انجام شده است. آزمایش‌هایی که در بخش ۴-۴ انجام شده است برای بررسی روش‌های خوشه‌بندی و مقایسه این روش‌ها با روش پیش‌نهادی در این رساله، انجام شده است.

در آزمایش انجام شده در بخش ۴-۴ میزان دقت خوشه‌بندی اسناد برای طرح‌های مختلف بازنمایی کلمات (چهار ماتریس هم‌رخدادی و V۲W)، استفاده از LDA و استفاده از ماتریس سند-کلمه (در دوحالت tf و tf*idf) مورد بررسی قرار گرفت. از طرفی برای هر یک از این حالات از الگوریتم‌های خوشه‌بندی کا-میانه، کامیانه-دوبخشی، خوشه‌بندی فازی و کانوپی استفاده شده است.

معیاری که برای ارزیابی خوشه بندی استفاده شده است معیار خلوص است. معیار خلوص یک خوشه نشان‌دهنده درصد درستی اسنادی است که به‌درستی در خوشه‌های خود قرار گرفته و به درستی خوشه‌بندی شده‌اند.

در ارزیابی اول با استفاده از خوشه‌بندی کامیانه روش‌های مختلف بازنمایی‌کلمات را مورد آزمون قرار داده‌ایم. همانطور که در جدول ۴-۱۹ ملاحظه می‌کنید در همه موارد روش ارائه‌شده در بازنمایی کلمات بسیار بهتر از روش‌های متداول در خوشه‌بندی اسناد (ماتریس‌های سند-کلمه) و روش LDA عمل می‌کند. عملکرد روش‌پیشنهادی با ۸ مجموعه‌داده مختلف مورد ارزیابی قرار گرفته و در میانگین ۷۵ درصد خلوص در خوشه‌بندی داشته است که اختلاف ۲۰ درصدی با روش‌های متداول خوشه‌بندی داشته است.

جدول ۴-۱۹ نشان می‌دهد که به طور کلی روش‌هایی که با استفاده از ماتریس هم‌رخدادی کلمات خوشه‌بندی سندها انجام شده است بهتر عملکرده‌اند. همچنین در جدول ۴-۱۹ نشان داده شده است که روش ارائه‌شده در نمایش هم‌رخدادی کلمات در مقایسه با دیگر روش‌های نمایش کلمات با میانگین ۷۵ درصد بهتر عمل کرده است. روش نمایش بیشترین هم‌رخدادی با میانگین ۷۲ درصد در جایگاه دوم و روش‌های شمارش سندها و تکرار کمتر در رتبه‌های بعدی قرار داشته‌اند (به ترتیب میانگین ۶۶ و ۵۹ درصد).

جدول ۴-۲۱ میزان خلوص خوشه‌بندی سندها را با استفاده از خوشه‌بندی کا-میانه-دوبخشی نشان می‌دهد. این جدول نیز همانند جدول ۴-۱۹ نشان می‌دهد که خوشه‌بندی با استفاده از روش‌های هم‌رخدادی کلمات از روش‌های متداول نمایش سند-کلمه (بر اساس فرکانس و tfidf) با اختلاف میانگین ۱۰ درصد عملکرد بهتری داشته‌اند. از میان روش‌های نمایش هم‌رخدادی کلمات، روش

نمایش بر حسب سند با میانگین ۷۲ و سپس روش پیشهادی با میانگین ۷۰ در رتبه دوم قرار گرفته است. روش‌های هم‌رخدادی حداکثر و حداقل نیز در رتبه‌های بعدی قرار دارند.

جدول ۴-۲۳ میزان خلوص در خوشه‌بندی اسناد را با استفاده از الگوریتم Canopy نشان می‌دهد. این خوشه‌بندی برای انواع روش‌های بازنمایی کلمات انجام شده است. همانطور که در این جدول ملاحظه می‌کنید استفاده از این روش خوشه‌بندی باعث افت شدید در دقت خوشه بندی شده است.

الگوریتم خوشه‌بندی دیگری که مورد آزمون قرار داده‌ایم، خوشه‌بندی فازی است. از آنجا که پارامتر فازی‌بودن (m) در کارایی الگوریتم تاثیر مستقیم دارد آزمون‌های متفاوتی را برای آن انجام داده‌ایم. در این آزمایش‌ها مقادیر فازی بودن بین ۱ تا ۲/۵ با فاصله‌های ۰/۲ قرار داده‌ایم و نتایج را برای روش‌های نمایش‌های مختلف هم‌رخدادی و روش‌های LDA و V2W مورد بررسی قرار داده‌ایم. نتایج در جداول ۲۵ تا ۴۲ نشان داده شده‌اند. با میانگین گرفتن نتایج خروجی این جداول برای هر یک از حالات خوشه‌بندی نتایج جدول ۴-۴۲ حاصل شد. این نتایج نشان می‌دهند که روش پیشنهادی برای نمایش هم‌رخدادی‌ها با استفاده از الگوریتم خوشه‌بندی خوشه‌بندی فازی با پارامتر m=۲.۱ از نظر دقت خوشه‌بندی کارایی بهتری از مابقی روش‌ها داشته است (۷۳/۵ ٪). الگوریتم خوشه‌بندی کا-میانه با میانگین ۷۱ ٪ در رتبه دوم و الگوریتم LDA با میانگین ۷۰ ٪ در رتبه سوم قرار دارند.

بنابراین آزمایش‌ها و ارزیابی‌های انجام شده نشان داد که روش ارائه در بازنمایی کلمات منجر به خوشه‌بندی با خلوص بهتری نسبت به خوشه‌بندی با استفاده از دیگر روش‌های بازنمایی کلمات بوده است. از طرفی دیگر استفاده از خوشه‌بندی فازی و کا-میانه نسبت به دیگر الگوریتم‌های خوشه‌بندی خلوص بیشتری را به همراه دارد و باعث بهبود در خوشه‌بندی اسناد می‌شود. از آنجا که در موضوع‌بندی عنوانی از خوشه‌بندی استفاده می‌شود توصیه می‌شود که برای این منظور از روش‌پیشنهادی به همراه روش خوشه‌بندی کا-میانه یا خوشه‌بندی فازی استفاده شود که در این رساله عملکرد بهتری نشان‌داده شده‌اند.

زمانی که هر سند بر اساس موضوع‌ها بازنمایی می‌شود، در اصل کلیدواژه‌هایی که در موضوع اصلی سند تاثیر چندانی ندارند به عنوان نویز حذف می‌شوند و مابقی کلیدواژه‌ها با تاثیر یکسانی در

موضوع سند تاثیر می‌گذارند. با استفاده از این نتیجه‌گیری (که برخی کلیدواژه‌ها در برخی اسناد نویز محسوب می‌شوند) می‌توانیم بگوییم که در مرحله تحلیل‌روند بهتر است تحلیل نمودار راهبردی بر اساس موضوع‌ها انجام شود تا تاثیر تکرار کلیدواژه‌هایی که به صورت نویز وجود دارند تعدیل شود.

از آنجا که با تبدیل ماتریس سند-کلمه به ماتریس سند-موضوع هر موضوع یک ویژگی محسوب می‌شود می‌توان گفت که به نوعی کاهش ویژگی صورت گرفته است. کاهش‌ویژگی انجام شده یک روش کاهش ویژگی غیر نظارت شده است. در پاراگراف‌های قبل به طور مفصل نشان دادیم که استفاده از این روش می‌تواند باعث افزایش میزان خلوص در خوشه‌بندی شود. اما در صورتی که از از روش ارائه‌شده برای استخراج ویژگی مطرح شده بخواهیم برای دسته‌بندی استفاده کنیم از پارامتر ارزیابی F۱ برای میزان صحت دسته‌بندی‌ها استفاده خواهد شد که در جدول ۴-۲۰ و ۲۳ آمده است. جدول ۴-۲۰ ملاحظه می‌کنید که میزان کاهش صحت در دسته‌بندی الگوریتم کاهش‌ویژگی ارائه‌شده با بهترین حالت ۴ درصد اختلاف دارد.

سهم علمی ارائه‌شده در این بخش ارائه یک روش نمایش هم‌رخدادی کلمات است که با استفاده از آن توانستیم خوشه‌بندی اسناد را به میزان قابل توجهی بهبود بخشیم. از طرفی در ادامه روشی با استفاده از هم‌رخدادی برای استخراج ویژگی برای اسناد نیز ارائه گردید. علاوه بر آن در مطالعات انجام شده روشی برای کشف موضوع‌ها با استفاده از روش خوشه‌بندی فازی یافت نشد که در این رساله نشان دادیم که استفاده از خوشه‌بندی فازی می‌تواند در مقادیر پایین m خوشه‌بندی‌های خوبی را ارائه دهد. سهم علمی دیگر این بخش بررسی روش‌های خوشه‌بندی بر روی اسناد علمی فارسی بوده است که در مطالعات گذشته یافت نشد.

۵-۳-۴- پرسش‌های پنجم: پیش‌بینی روند تحقیقاتی بر اساس رساله‌های انجام‌شده به چه صورت خواهد بود؟

با استفاده از روش‌های دسته بندی می‌توان بعد از تفکیک حوزه‌های علمی، روند حرکت آن‌ها را در سال‌های متمادی بررسی کرد. همانطور که در شکل‌های ۳۳ و ۳۴ ملاحظه می‌کنید، بر اساس مجموعه داده‌های دریافت شده از پایگاه گنج، پارسای موجود در سال‌های ۹۴ نسبت به سال‌های

گذشته بسیار افت کرده است. این روند در میزان پارسای ثبت شده از طرف دانشگاه‌های نیز دیده شده است.شکل ۴-۱۸ برخی دانشگاه‌هایی که بیشترین پژوهش‌ها را در حوزه برق در سامانه ثبت کرده‌اند را نشان می‌دهد که این کاهش را در آن نیز مشاهده می‌کنید. *لازم به ذکر است که در این رساله به بررسی دلیل این کاهش پرداخته نمی‌شود و تنها روند کشف شده را به نمایش گذاشته می‌شود.*

با استفاده از روشی که در بخش ۳-۴-۳- بیان شد، تعداد موضوع‌هایی که در آن پارامترهای خوشه‌بندی بهینه‌تر شده باشند انتخاب می‌شوند. بر اساس تعداد خوشه‌ای که در مرحله قبل بدست آمد حوزه مورد نظر را با استفاده از الگوریتم پیشنهادی توسط یکی از الگوریتم‌های خوشه‌بندی کا-میانه یا خوشه‌بندی فازی خوشه‌بندی خواهیم کردیم که تعداد ۱۲ خوشه درست شدند. کلمات پر تکرار هر یک از این خوشه‌ها را در جدول ۴-۶۵ آورده شده است. همانطور که در شکل ۳-۱ ملاحظه می‌شود، شخص تحلیل‌گر می‌تواند بر اساس تشخیص خود خوشه‌ها را نظارت کند و خوشه‌هایی که به صورت کلی بیان شده‌اند و موضوع خاصی را اشاره نمی‌کنند را حذف کند. برای مثال موضوع شماره ۴ دارای یک کلمه است و موضوع خاصی را اشاره نمی‌کند. در ادامه شخص یا اشخاص متخصص برای سادگی کار می‌توانند بر اساس نظر خود بر روی هر یک از موضوع‌ها برچسب بزنند.

بر این اساس موضوع‌های تشخیص داده شده جدول ۴-۷۱ تا جدول ۴-۷۸ طراحی شده‌اند. این نشان می‌دهد می‌توان به **سوال پنجم** پژوهش در خصوص پیش‌بینی روند پژوهش پاسخ داد. این جدول نشان می‌دهد که روند رشد هر موضوع در سال‌های گذشته به چه صورت بوده است.

بعد از استخراج موضوع‌ها می‌توان میزان تحقیقات دانشگاه‌های مختلف را بر روی موضوع‌های متختلف بررسی کرد. مجدول ۴-۶۶ متوسط سهم موضوع‌های انجام‌شده در یک رساله در دانشگاه‌های با بیشترین پارساهای ثبت‌شده (برای مقایسه پذیری اعداد داخل جدول برحسب تعداد پارساها نرمال شده‌اند و در ۱۰۰ ضرب شده) را نشان می‌دهد. این جدول نشان می‌دهد که هر دانشگاه

به طور متوسط در هر پارسا بر روی چند موضوع فعالیت کرده است و بیشترین فعالیت هر دانشگاه بر روی کدوم موضوع، یا کدام موضوع‌ها بوده است. برای مثال دانشگاه بیرجند بیشتر از دانشگاه‌های دیگر بر روی موضوع ۱۱ کار می‌کند. دانشگاه شاهرود بر روی موضوع ۵ بیشتر و دانشگاه تبریز بر روی موضوع ۲ تمرکز بیشتری دارد.

۵-۴- کاربردها

تمام شرکت‌ها و سازمان‌های بزرگ و کوچک برای رسیدن به اهداف و چشم‌اندازهای کوتاه‌مدت و بلندمدت خود، برنامه‌ریزی‌هایی انجام می‌دهند. لازمه انجام چنین برنامه‌ریزی‌هایی تحلیل‌روند وضعیت فعلی و پیش‌بینی آینده آن حوزه از فعالیت‌ها است. برای مثال نقشه جامع علمی کشور ازجمله نمونه‌های این نوع تحلیل‌ها است. تحلیل روند شامل مجموعه ای از اطلاعات مرتبط با بازه های زمانی مختلف جهت بازنگری های متنوع در سیاست های سازمانی است.

بنابراین با استفاده از روش ارائه‌شده در این رساله علاوه تحلیل‌روند در سطح کلان، محققان می‌توانند از حوزه مورد مطالعه خود دیدی کامل و جامع داشته باشند. از طرفی با استفاده از اجزای روش ارائه‌شده تحلیل‌های نمودار راهبردی می‌تواند در سطحی بالاتر از کلمات انجام شود. با استفاده از روش ارائه‌شده در خصوص بازنمایی کلمات نیز می‌تواند در تفکیک غیرنظارت شده حوزه‌های علمی به صورت موثری استفاده شود.

۵-۵- محدودیت‌های پژوهش

در زیر محدودیت‌هایی که در این پژوهش وجود داشته را بر شمرده و برای برخی از آن‌ها پیشنهادهایی برای کارهای آینده بیان کرده‌ایم :

- یکی از بزرگترین محدودیت‌هایی که در پژوهش با آن روبرو هستیم تاثیر مستقیم سیاست‌های انتخاب شده در بازه‌های زمانی متفاوت بر روی کمیت و کیفیت اسناد علمی است که در کشور انجام می‌شود. تغییرات فراوان در مدیریت و اتخاذ سیاست‌های متفاوت به صورت مستقیم در کیفیت خروجی تاثیر خواهد گذاشت.

پیشنهاد می‌شود علاوه بر سیستم خودکاری که از اسناد علمی استفاده می‌کند از سیستمی با پارامترهای اقتصادی و سیاسی کمک گرفته شود.

- محدودیت در نوع سندهای علمی در دسترس: همانطور که پیشینه‌پژوهش به آن اشاره شد، از اسناد علمی متفاوت می‌توان در تحلیل‌روند استفاده کرد. اسناد علمی می‌توانند جنبه‌ها مختلف از یک حوزه را به نمایش بگذارند. بسیاری از تحلیل‌ها و پیش‌بینی‌های فناوری با استفاده از ثبت اختراعات انجام شده است. از آنجا که در این پژوهش تنها از پارسا استفاده شده است، برای پژوهش‌های آینده و افزایش دقت در تحلیل‌های روند می‌توان از اسناد علمی دیگری، همچون مقاله‌ها و اختراعات ثبت شده در آن حوزه بهره گرفت.
- محدودیت در تعداد و زمان پارسا: برای تحلیل‌روند با دقت مناسب لازم است که اسناد مورد بررسی از نظر تعداد بسیار زیاد و به صورت جامع از تمام دانشگاه‌ها تهیه شده باشند. از آنجا که در انتخاب اسناد موجود محدودیت‌های زمانی و تعداد وجود داشت این محدودیت می‌تواند در میزان دقت در نتیجه تاثیراتی برجا بگذارد.
- خطای ابزارهای مورد استفاده متن‌کاوی (نرمال سازی، برچسب‌زنی کلمات): همه ابزارهای موجود دارای محدودیت‌هایی هستند که می‌توانند موجب خطاهایی بروز خطا شوند. ابزار مورد استفاده در این رساله توسط آزمایشگاه دانشگاه فردوسی درست شده است. در هنگام استفاده از این ابزار در هنگام نرمال کردن متن با خطاهایی پیش‌بینی نشده و نا معلومی مواجه بودیم که برای از دست ندادن متن مجبور به استفاده از آن متن به صورت نرمال نشده داشتیم. در حالاتی نیز ابزار ریشه‌یابی این برخی کلمات را به درستی ریشه‌یابی نکرده بود که به صورت دستی آن کلمات تغییر و یا حذف شدند. به همین دلیل بخشی برای بررسی مجدد توسط متخصصین در این رساله اضافه شده است.
- نبود کلمات‌ایست جامع یک حوزه علمی: وجود کلمات ایستی که قبلا توسط متخصص‌ها ارائه شده‌اند علاوه بر اینکه سرعت سیستم را بالا می‌برد، در دقت نتایج

تاثیر مستقیم می‌گذارد. در صورتی که چند کلمه‌ایست به عنوان کلیدواژه به مرحله تولید ماتریس‌های هم‌رخدادی و موضوع بندی بشوند می‌توانند تاثیرات نامناسبی به نتایج خوشه‌بندی وارد کنند. بنابراین پیشنهادی می‌شود که تحلیل‌روند را با استفاده از لیست کلمات‌ایست که توسط متخصصان آماده شده است اجرا شود. برای این منظور می‌توان روشی ارائه کرد که با اسناد بسیار زیادی از یک حوزه این لیست را استخراج و از آن استفاده نمایند.

- عدم تطابق روابط معنایی کلمات چکیده با اصطلاح نامه: برای بهبود در خوشه‌بندی علاوه بر اینکه از پارامترهای ارزیابی می‌توان استفاده کرد، از اصطلاح‌نامه‌ها یا آنتولوژی‌ها می‌توان استفاده نمود. از آنجا که در پژوهشی که انجام شد مشخص شد که روابط معنایی کلمات کلیدواژه‌ها با اصطلاح‌نامه‌ها تطابق ندارند استفاده از اصطلاح‌نامه‌های موجود در این پژوهش نا ممکن شد. پیشنهاد می‌شود با استفاده از روش‌هایی که روابط معنایی کلمات را در متن استخراج می‌کنند اصطلاح‌نامه‌ای برای حوزه‌ها تشکیل داد و از آن برای خوشه‌بندی معنایی کلمات به صورت دقیق‌تر در تحلیل‌روند استفاده نمود.
- تشخیص تعداد خوشه‌ها یکی از زمینه‌های مهم تحقیقاتی در یادگیری‌های غیرنظارت شده در داده‌کاوی است که هنوز راه حل دقیقی برای آن ارائه نشده است. یکی از چالش‌های مهم موجود در این رساله تعیین تعداد خوشه‌ها بوده‌است. تعیین تعداد خوشه‌ها می‌تواند در نتیجه تاثیر مستقیم قرار دهد و از آنجا که استفاده از تعدادی معیار ارزیابی خوشه به تنهایی نمی‌توانند بهترین ارزیابی را به ما بدهند، پیشنهاد می‌شود که برای تحلیل‌های قوی‌تر پژوهشی به صورت جداگانه بر روی نحوه ارزیابی متون علمی در یک خوشه‌بندی انجام شود.

۵-۶- پیشنهاد برای کارهای آتی

همانطور که بیان شد تغییرات فراوان در مدیریت و اتخاذ سیاست‌های متفاوت به صورت مستقیم در کیفیت خروجی تاثیر خواهد گذاشت. پیشنهاد می‌شود علاوه بر سیستم خودکاری که از اسناد علمی استفاده می‌کند از سیستمی با پارامترهای اقتصادی و سیاسی کمک گرفته شود. با ترکیب این پارامترها و نظرات متخصصان می‌توان پیش‌بینی و تحلیل دقیق‌تری از روند علم و فناوری کشور بدست آورد.

یکی از روش‌های خوشه‌بندی استفاده از اصطلاحنامه‌ها و منابعی است که بتوانند ارتباط معنایی دقیق‌تری را نشان‌دهند. از آنجا که در پژوهش‌های گذشته نشان داده شد که اصطلاحنامه‌های عمومی در اسناد علمی کاربرد زیادی ندارند پیشنهاد می‌شود در پژوهش‌های آتی از ابزارهای که این ارتباط معنایی را به صورت دقیق‌تری مشخص می‌کنند استفاده شود. از طرفی در صورت زیاد بودن تعداد اسناد می‌توان با استفاده از ابزارهای دیگری این روبط معنایی را در متون استخراج کرد و از آن‌ها استفاده کرد.

از طرفی دیگر برای جامع بودن تحلیل‌روند می‌توان سیستمی ارائه کرد که ترکیبی از اسناد مختلف از جمله اختراعات ثبت شده، روزنامه‌ها و سایت‌های فناوری استفاده کند. چنین پژوهشی نیاز به بررسی ماهیت داده‌ها در هر نوع و وزن‌دهی مناسب از اطلاعات مختلف است.

مراجع

Abilhoa, Willyan D, and Leandro N de Castro. ۲۰۱۴. "A keyword extraction method from twitter messages represented as graphs." *Applied Mathematics and Computation* ۲۴۰:۳۰۸-.۳۲۵

Aggarwal, Charu C, and ChengXiang Zhai. ۲۰۱۲. "A survey of text clustering algorithms." In *Mining text data*, ۷۷-۱۲۸. Springer.

Aghdam, Mehdi Hosseinzadeh, Nasser Ghasem-Aghaee, and Mohammad Ehsan Basiri. ۲۰۰۹. "Text feature selection using ant colony optimization." *Expert Systems with Applications* ۳۶ (۳):۶۸۴۳-.۶۸۵۳

AGUS WIDODO, NOVITASARI NAOMI, SUHARJITO. ۲۰۱۳. "Prediction of research topics using combination of machine learning and logistic curve." *Journal of Theoretical and Applied Information Technology* .(۳) ۴۹

Aha, David W. ۱۹۹۸. "Feature weighting for lazy learning algorithms." In *Feature Extraction, Construction and Selection*, ۱۳-۳۲. Springer.

Akogul, Serkan, and Murat Erisoglu. ۲۰۱۷. "An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis." *Entropy* ۱۹ (۹):.۴۵۲

Alelyani, Salem, Jiliang Tang, and Huan Liu. ۲۰۱۳. "Feature Selection for Clustering: A Review." *Data Clustering: Algorithms and Applications* .۲۹

Altszyler, Edgar, Mariano Sigman, Sidarta Ribeiro, and Diego Fernández Slezak. ۲۰۱۶. "Comparative study of LSA vs Word۲vec embeddings in small corpora: a case study in dreams database." *arXiv preprint arXiv:*۱۶۱۰..۰۱۵۲۰

An, Xin Ying, and Qing Qiang Wu. ۲۰۱۱. "Co-word analysis of the trends in stem cells field based on subject heading weighting." *Scientometrics* ۸۸ (۱):۱۳۳-.۱۴۴

Anna Sokolova ,Nadezhda Mikova. ۲۰۱۴. "Comparing information sources for identifying technology trends." Comparing information sources for identifying technology trends, Brussels, ۲۷ November.

Ansari, Mariam. ۲۰۰۵. "Matching between assigned descriptors and title keywords in medical theses." *Library Review* ۵۴ (۷):۴۱۰-.۴۱۴

Baharudin, Baharum, Lam Hong Lee, and Khairullah Khan. ۲۰۱۰. "A review of machine learning algorithms for text-documents classification." *Journal of advances in information technology* ۱ (۱):۴-.۲۰

Banerjee, Shreya, Ankit Choudhary, and Somnath Pal. ۲۰۱۵. "Empirical evaluation of k-means, bisecting k-means, fuzzy c-means and genetic k-means clustering algorithms." Electrical and Computer Engineering (WIECON-ECE), ۲۰۱۵ IEEE International WIE Conference on.

Beil, Florian, Martin Ester, and Xiaowei Xu. ۲۰۰۲. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. ۲۰۰۳. "A neural probabilistic language model." *Journal of machine learning research* ۳ (Feb):۱۱۳۷-.۱۱۵۵

Bengisu, Murat, and Ramzi Nekhili. ۲۰۰۶. "Forecasting emerging technologies with the aid of science and technology databases." *Technological Forecasting and Social Change* ۷۳ (۷):۸۳۵-۸۴۴

Bian, Jiang, Bin Gao, and Tie-Yan Liu. ۲۰۱۴. "Knowledge-powered deep learning for word embedding." Joint European conference on machine learning and knowledge discovery in databases.

Blei, David M, Andrew Y Ng, and Michael I Jordan. ۲۰۰۳. "Latent dirichlet allocation." *Journal of machine Learning research* ۳ (Jan):۹۹۳-.۱۰۲۲

Buscaino, B. ۲۰۱۴. "Trend Analysis in Biomedical Texts via Vector Space Model Synonym Extraction." *Technická zpráva, Lawrence Livermore National Laboratory (LLNL), Livermore, CA*.

Bush, Vannevar. ۱۹۴۵. "Science: The endless frontier." *Transactions of the Kansas Academy of Science (۱۹۰۳)*:۲۳۱-.۲۶۴

Callon, Michel, Jean-Pierre Courtial, William A Turner, and Serge Bauin. ۱۹۸۳. "From translations to problematic networks: An introduction to co-word analysis." *Information (International Social Science Council)* ۲۲ (۲):۱۹۱-.۲۳۵

Callon, Michel, Jean Pierre Courtial, and Francoise Laville. ۱۹۹۱. "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry." *Scientometrics* ۲۲ (۱):۱۵۵-.۲۰۵

Campr, Michal, and Karel Ježek. ۲۰۱۵. "Comparing semantic models for evaluating automatic document summarization." International Conference on Text, Speech, and Dialogue.

Carley, Stephen F, Nils C Newman, Alan L Porter, and Jon G Garner. ۲۰۱۷. "A measure of staying power: Is the persistence of emergent concepts more significantly influenced by technical domain or scale?" *Scientometrics* ۱۱۱ (۳):۲۰۷۷-.۲۰۸۷

Carley, Stephen F, Nils C Newman, Alan L Porter, and Jon G Garner. ۲۰۱۸. "An indicator of technical emergence." *Scientometrics* ۱۱۵ (۱):۳۵-.۴۹

Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella. ۲۰۱۰. "Emerging topic detection on twitter based on temporal and social terms evaluation." Proceedings of the tenth international workshop on multimedia data mining.

Chandrashekar, Girish, and Ferat Sahin. ۲۰۱۴. "A survey on feature selection methods." *Computers & Electrical Engineering* ۴۰ (۱):۱۶-.۲۸

Chang, Xing, Xin Zhou, Linzhi Luo, Chengjia Yang, Hui Pan, and Shuyang Zhang. ۲۰۱۷. "Hotspots in research on the measurement of medical students' clinical competence from ۲۰۱۲-۲۰۱۶ based on co-word analysis." *BMC medical education* ۱۷ (۱):۱۶۲.

Chang, Yueh-Hsia, Chun-Yen Chang, and Yuen-Hsien Tseng. ۲۰۱۰. "Trends of science education research: An automatic content analysis." *Journal of Science Education and Technology* ۱۹ (۴):۳۱۵-.۳۳۱

Chao, Chia-Chen, Jiann-Min Yang, and Wen-Yuan Jen. " .۲۰۰۷Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from ۱۹۹۱ to ۲۰۰۵." *Technovation* ۲۷ (۵):۲۶۸-.۲۷۹

Chen, Jingnian, Houkuan Huang, Shengfeng Tian, and Youli Qu. ۲۰۰۹. "Feature selection for text classification with Naïve Bayes." *Expert Systems with Applications* ۳۶ (۳):۵۴۳۲-.۵۴۳۵

Chen, Xiuwen, Jianming Chen, Dengsheng Wu, Yongjia Xie, and Jing Li. ۲۰۱۶. "Mapping the research trends by co-word analysis based on keywords from funded project." *Procedia Computer Science* ۹۱:۵۴۷-.۵۵۵

Cheng, Xueqi, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. ۲۰۱۴. "Btm: Topic modeling over short texts." *IEEE Transactions on Knowledge & Data Engineering* (۱):۱-.۱

Chiu, Billy, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. " .۲۰۱۶How to train good word embeddings for biomedical NLP." Proceedings of the ۱۵th Workshop on Biomedical Natural Language Processing.

Choi, Changwoo, and Yongtae Park. ۲۰۰۹. "Monitoring the organic structure of technology based on the patent development paths." *Technological Forecasting and Social Change* ۷۶ (۶):۷۵۴-.۷۶۸

Christodoulos, Charisios, Christos Michalakelis, and Dimitris Varoutas. ۲۰۱۰. "Forecasting with limited data: Combining ARIMA and diffusion models." *Technological Forecasting and Social Change* ۷۷ (۴):۵۵۸-.۵۶۵

Coates, Vary, Mahmud Farooque, Richard Klavans, Koty Lapid, Harold A Linstone, Carl Pistorius, and Alan L Porter. ۲۰۰۱. "On the future of technological forecasting." *Technological Forecasting and Social Change* ۶۷ (۱):۱-.۱۷

Cover ,Thomas, and Peter Hart. ۱۹۶۷. "Nearest neighbor pattern classification." *IEEE transactions on information theory* ۱۳ (۱):۲۱-.۲۷

Cover, Thomas M, and Joy A Thomas. ۲۰۱۲. *Elements of information theory*: John Wiley & Sons.

Cozzens, Susan, Sonia Gatchair, Jongseok Kang, Kyung-Sup Kim, Hyuck Jai Lee, Gonzalo Ordóñez, and Alan Porter. ۲۰۱۰. "Emerging technologies:

quantitative identification and measurement." *Technology Analysis & Strategic Management* ۲۲ (۳):۳۶۱-.۳۷۶

Daim, Tugml U, Guillermo R Rueda, and Hilary T Martin. ۲۰۰۵. "Technology forecasting using bibliometric analysis and system dynamics." Technology management: a unifying discipline for melting the boundaries.

Daim, Tugrul U, Guillermo Rueda, Hilary Martin, and Pisek Gerdsri. ۲۰۰۶. "Forecasting emerging technologies: Use of bibliometrics and patent analysis." *Technological Forecasting and Social Change* ۷۳ (۸):۹۸۱-.۱۰۱۲

Das, Sanmay. ۲۰۰۱. "Filters, wrappers and a boosting-based hybrid for feature selection." ICML.

Dash, Manoranjan, and Poon Wei Koot. ۲۰۰۹. "Feature selection for clustering." In *Encyclopedia of database systems*, ۱۱۱۹-۱۱۲۵. Springer.

Davarpanah, Mohammad Reza, M Sanji, and M Aramideh. ۲۰۰۹. "Farsi lexical analysis and stop word list." *Library Hi Tech* ۲۷ (۳):۴۳۵-.۴۴۹

Dawelbait, Gihan ,Andreas Henschel, Toufic Mezher, and Wei Lee Woon. ۲۰۱۱. "Forecasting Renewable Energy Technologies in Desalination and Power Generation Using Taxonomies." *International Journal of Social Ecology and Sustainable Development (IJSESD)* ۲ (۳):۷۹-.۹۳

Dawelbait, Gihan, Toufic Mezher, Wei Lee Woon, and Andreas Henschel. ۲۰۱۰. "Taxonomy based trend discovery of renewable energy technologies in desalination and power generation." Technology Management for Global Economic Growth (PICMET), ۲۰۱۰ Proceedings of PICMET'.:۱۰

de Amorim, Renato Cordeiro, and Christian Hennig. ۲۰۱۵. "Recovering the number of clusters in data sets with noise features using feature rescaling factors." *Information Sciences* ۳۲۴:۱۲۶-.۱۴۵

de la Hoz-Correa, Andrea, Francisco Muñoz-Leiva, and Márta Bakucz. ۲۰۱۸. "Past themes and future trends in medical tourism research: A co-word analysis." *Tourism Management* ۶۵:۲۰۰-.۲۱۱

Dehdarirad, Tahereh, Anna Villarroya, and Maite Barrios. ۲۰۱۴. "Research trends in gender differences in higher education and science: a co-word analysis." *Scientometrics* ۱۰۱ (۱):۲۷۳-.۲۹۰

Delecroix, Bertrand, and R Epstein. ۲۰۰۴. "Co-word analysis for the non-scientific information example of Reuters Business Briefings." *Data Science Journal* ۳:۸۰-۸۷

Ercan, Gonenc, and Ilyas Cicekli. ۲۰۰۷. "Using lexical chains for keyword extraction." *Information processing & management* ۴۳ (۶):۱۷۰۵-۱۷۱۴

Esch, Maurice E. ۱۹۶۵. "Planning assistance through technical evaluation of relevance numbers(Planning Assistance Through Technical Evaluation of Relevance Numbers/PATTERN/ DISCUSSED as aid to decision-making in military technology area)." NATIONAL AEROSPACE ELECTRONICS CONFERENCE, ۱۷ TH, DAYTON, OHIO.

Fahad, Adil, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. ۲۰۱۴. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." *IEEE transactions on emerging topics in computing* ۲ (۳):۲۶۷-۲۷۹

Fattori, Michele, Giorgio Pedrazzi, and Roberta Turra. " .۲۰۰۳Text mining applied to patent mapping: a practical business case." *World Patent Information* ۲۵ (۴):۳۳۵-۳۴۲

Fayyad, Usama M, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. ۱۹۹۶. *Advances in knowledge discovery and data mining*. Vol. ۲۱: AAAI press Menlo Park.

Fayyd, Usama M, Gregory P Shapiro, and Padhraic Smyth. ۱۹۹۶. "From data mining to knowledge discovery: an overview".

Firat, Ayse Kaya, Wei Lee Woon, and Stuart Madnick. ۲۰۰۸. "Technological forecasting–A review." *Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology*.

Firth, John R. ۱۹۵۷. "A synopsis of linguistic theory, ۱۹۳۰-۱۹۵۵." *Studies in linguistic analysis*.

Fiscus, Jonathan G, and George R Doddington. ۲۰۰۲. "Topic detection and tracking evaluation overview." In *Topic detection and tracking*, ۱۷-۳۱. Springer.

Fodeh, Samah, Bill Punch, and Pang-Ning Tan. ۲۰۱۱. "On ontology-driven document clustering using core semantic features." *Knowledge and information systems* ۲۸ (۲):۳۹۵-.۴۲۱

Ghanem ,Moustafa M, Yike Guo, Huma Lodhi, and Yong Zhang. ۲۰۰۲. "Automatic scientific text classification using local patterns: KDD Cup ۲۰۰۲ (task ۱)." *ACM SIGKDD Explorations Newsletter* ۴ (۲):۹۵-.۹۶

Ghani, Rayid, and Andrew E Fano. ۲۰۰۲. "Using text mining to infer semantic attributes for retail data mining." Data Mining, ۲۰۰۲. ICDM ۲۰۰۳. Proceedings. ۲۰۰۲ IEEE International Conference on.

Glänzel, Wolfgang, and Bart Thijs. ۲۰۱۱. "Using 'core documents' for detecting and labelling new emerging topics." *Scientometrics* ۹۱ (۲):۳۹۹-.۴۱۶

Gordon, Theodore J, and Olaf Helmer. ۱۹۶۴. Report on a long-range forecasting study. Rand Corporation Santa Monica, CA.

Guo, Daoyan, Hong Chen, Ruyin Long, Hui Lu, and Qianyi Long. ۲۰۱۷. "A co-word analysis of organizational constraints for maintaining sustainability." *Sustainability* ۹ (۱۰):.۱۹۲۸

Guyon, Isabelle, and André Elisseeff. ۲۰۰۳. "An introduction to variable and feature selection." *Journal of machine learning research* ۳ (Mar):۱۱۵۷-.۱۱۸۲

Habibi, Maryam, and Andrei Popescu-Belis. ۲۰۱۵. "Keyword extraction and clustering for document recommendation in conversations." *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* ۲۳ (۴):۷۴۶-.۷۵۹

Halaweh, Mohanad. ۲۰۱۳. "Emerging technology: What is it." *Journal of technology management & innovation* ۸ (۳):۱۰۸-.۱۱۵

Han, Jiawei, Jian Pei, and Micheline Kamber. ۲۰۱۱. *Data mining: concepts and techniques*: Elsevier.

Harris, Zellig S. ۱۹۵۴. "Distributional structure." *Word* ۱۰ (۲-۳):۱۴۶-.۱۶۲

He, Qin. ۱۹۹۹. "Knowledge discovery through co-word analysis." *Library trends* ۴۸ (۱):۱۳۳-.۱۳۳

Henschel, Andreas, Erik Casagrande, Wei Lee Woon, Isam Janajreh, and Stuart Madnick. ۲۰۱۲. "A Unified Approach for Taxonomy-Based Technology Forecasting." *Business Intelligence Applications and the Web: Models, Systems and Technologies*:.۱۷۸

Henschel, Andreas, Wei Lee Woon, Thomas Wachter, and Stuart Madnick. ۲۰۰۹. "Comparison of generality based algorithm variants for automatic taxonomy generation." Innovations in Information Technology .۲۰۰۹ ,IIT'۰۹. International Conference on.

Hong, Bao, and Deng Zhen. ۲۰۱۲. "An extended keyword extraction method." *Physics Procedia* ۲۴:۱۱۲۰-.۱۱۲۷

Hotho, Andreas, Steffen Staab, and Gerd Stumme. ۲۰۰۳. "Ontologies improve text document clustering." Data Mining, ۲۰۰۳. ICDM ۲۰۰۳. Third IEEE International Conference on.

Hu, Chang-Ping, Ji-Ming Hu, Sheng-Li Deng, and Yong Liu. ۲۰۱۳. "A co-word analysis of library and information science in China." *Scientometrics* ۹۷ (۲):۳۶۹-.۳۸۲

Hu, Jian, Lujun Fang, Yang Cao ,Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. ۲۰۰۸. "Enhancing text clustering by leveraging Wikipedia semantics." Proceedings of the ۳۱st annual international ACM SIGIR conference on Research and development in information retrieval.

Hu, Jiming, and Yin Zhang. ۲۰۱۵. "Research patterns and trends of Recommendation System in China using co-word analysis." *Information processing & management* ۵۱ (۴):۳۲۹-.۳۳۹

Hulth, Anette. ۲۰۰۳. "Improved automatic keyword extraction given more linguistic knowledge ".Proceedings of the ۲۰۰۳ conference on Empirical methods in natural language processing.

Jain, Anil K. ۲۰۱۰. "Data clustering: ۵۰ years beyond K-means." *Pattern recognition letters* ۳۱ (۸):۶۵۱-.۶۶۶

Jain, Anil K, and Richard C Dubes. ۱۹۸۸. "Algorithms for clustering data".

Jain, Anil K, M Narasimha Murty, and Patrick J Flynn. ۱۹۹۹. "Data clustering: a review." *ACM computing surveys (CSUR)* ۳۱ (۳):۲۶۴-.۳۲۳

Jain, Anil, and Douglas Zongker. ۱۹۹۷. "Feature selection: Evaluation, application, and small sample performance." *IEEE transactions on pattern analysis and machine intelligence* ۱۹ (۲):۱۵۳-.۱۵۸

Jalalimanesh, Ammar. ۲۰۱۲. "Knowledge discovery in scientific databases using text mining and social network analysis." Control, Systems & Industrial Informatics) ICCSII), ۲۰۱۲ IEEE Conference on.

Jiang, Shenhao, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. ۲۰۱۸. "Identifying Emergent Research Trends by Key Authors and Phrases." Proceedings of the ۲۷th International Conference on Computational Linguistics.

Jo, Yookyung, Carl Lagoze, and C Lee Giles. ۲۰۰۷. "Detecting research topics via the correlation between graphs and texts." Proceedings of the ۱۳th ACM SIGKDD international conference on Knowledge discovery and data mining.

Joung, Junegak, and Kwangsoo Kim" .۲۰۱۷ .Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data." *Technological Forecasting and Social Change* ۱۱۴:۲۸۱-.۲۹۲

Jun, Sunghae. ۲۰۱۱a. "A forecasting model for technological trend using unsupervised learning." In *Database Theory and Application, Bio-Science and Bio-Technology*, ۵۱-۶۰. Springer.

Jun, Sunghae. ۲۰۱۱b. "IPC Code Analysis of Patent Documents Using Association Rules and Maps–Patent Analysis of Database Technology." In *Database Theory and Application, Bio-Science and Bio-Technology*, ۲۱-۳۰. Springer.

Jun, Sunghae, and Seung-Joo Lee. ۲۰۱۲. "Emerging Technology Forecasting Using New Patent Information Analysis." *International Journal of Software Engineering & Its Applications* .(۳) ۶

Jun, Sunghae, Sang Sung Park, and Dong Sik Jang. ۲۰۱۲. "Technology forecasting using matrix map and patent clustering." *Industrial Management & Data Systems* ۱۱۲ (۵):۷۸۶-.۸۰۷

Karami, Amir, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. ۲۰۱۵. "A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections." *iConference ۲۰۱۵ Proceedings*.

Karypis, Michael Steinbach George, Vipin Kumar, and Michael Steinbach. ۲۰۰۰. "A comparison of document clustering techniques." KDD workshop on Text Mining.

Kasiviswanathan, Shiva Prasad, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. ۲۰۱۱. "Emerging topic detection using dictionary learning." Proceedings of the ۲۰th ACM international conference on Information and knowledge management.

Kaufman, Leonard, and Peter J Rousseeuw. ۲۰۰۹. *Finding groups in data: an introduction to cluster analysis*. Vol. ۳۴۴: John Wiley & Sons.

Khatir, Ashkan, and Soheil Ganjefar. ۲۰۱۶. "The Analysis of the Distribution and Focus of Keywords in Theses and Dissertations: The Compliance with Descriptors, Title, and Abstract." *Journal of Information Processing and Management*.

Khazaei, A, and M Ghasemzadeh. ۲۰۱۵. "Comparing k-means clusters on parallel Persian-English corpus." *Journal of AI and Data Mining* ۳ (۲):۲۰۳-.۲۰۸

Kim, Mee-Jean. ۲۰۰۷. "A bibliometric analysis of the effectiveness of Korea's Biotechnology Stimulation Plans, with a comparison with four other Asian nations." *Scientometrics* ۷۲ (۳):۳۷۱-.۳۸۸

Kim, Yoosin, Yeonjin Ju, SeongGwan Hong, and Seung Ryul Jeong. ۲۰۱۷. "Practical Text Mining for Trend Analysis: Ontology to visualization in Aerospace Technology." *KSII Transactions on Internet & Information Systems* .(۸) ۱۱

Kim, Young Gil, Jong Hwan Suh, and Sang Chan Park. ۲۰۰۸. "Visualization of patent analysis for emerging technology." *Expert Systems with Applications* ۳۴ (۳):۱۸۰۴-.۱۸۱۲

Kivikunnas, Sauli. ۱۹۹۸. "Overview of process trend analysis methods and applications." ERUDIT Workshop on Applications in Pulp and Paper Industry.

Klavans, Richard, and Kevin W Boyack. ۲۰۱۷. "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?" *Journal of the Association for Information Science and Technology* ۶۸ (۴):۹۸۴-.۹۹۸

Kobayashi, Shin-ichi, Yasuyuki Shirai, Kazuo Hiyane, Fumihiro Kumeno, Hiroshi Inujima, and Noriyoshi Yamauchi. ۲۰۰۵. "Technology trends analysis from the internet resources." In *Advances in Knowledge Discovery and Data Mining*, ۸۲۰-۸۲۵. Springer.

Kohavi, Ron, and George H John. " .۱۹۹۷Wrappers for feature subset selection." *Artificial intelligence* ۹۷ (۱):۲۷۳-.۳۲۴

Kou, Gang, Yi Peng, and Guoxun Wang. ۲۰۱۴. "Evaluation of clustering algorithms for financial risk analysis using MCDM methods." *Information Sciences* ۲۷۵:۱-.۱۲

Kovács, Ferenc, Csaba Legány, and Attila Babos. ۲۰۰۵. "Cluster validity measurement techniques." ۶th International symposium of hungarian researchers on computational intelligence.

Kung, Yen-Ying, Shinn-Jang Hwang, Tsai-Feng Li, Seong-Gyu Ko, Ching-Wen Huang, and Fang–Pey Chen. ۲۰۱۷. "Trends in global acupuncture publications: An analysis of the Web of Science database from ۱۹۸۸ to ۲۰۱۵." *Journal of the Chinese Medical Association* ۸۰ (۸):۵۲۱-.۵۲۵

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. ۲۰۱۵. "From word embeddings to document distances." International Conference on Machine Learning.

Lai, Siwei, Kang Liu, Shizhu He, and Jun Zhao. ۲۰۱۶. "How to generate a good word embedding." *IEEE Intelligent Systems* ۳۱ (۶):۵-.۱۴

Larsen, Bjornar, and Chinatsu Aone. ۱۹۹۹. "Fast and effective text mining using linear-time document clustering." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining.

Lau, Jey Han, Nigel Collier, and Timothy Baldwin. ۲۰۱۲. "On-line trend analysis with topic models:\# twitter trends detection topic model online." *Proceedings of COLING* ۲۰۱۲:۱۵۱۹-.۱۵۳۴

Law, John, Serge Bauin, J Courtial, and John Whittaker. ۱۹۸۸. "Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification." *Scientometrics* ۱۴ (۳-۴):۲۵۱-.۲۶۴

Lee, Li-Wei, and Shyi-Ming Chen. ۲۰۰۶. "New methods for text categorization based on a new feature selection method and a new similarity measure between documents." In *Advances in Applied Artificial Intelligence*, ۱۲۸۰-۱۲۸۹. Springer.

Lee, Sungjoo, Byungun Yoon, and Yongtae Park. ۲۰۰۹. "An approach to discovering new technology opportunities: Keyword-based patent map approach." *Technovation* ۲۹ (۶):۴۸۱-.۴۹۷

Levary, Reuven R, and Dongchui Han" .۱۹۹۵ .Choosing a technological forecasting method." *Industrial Management* ۳۷ (۱):۱۴-.۱۸

Leydesdorff, Loet, and Adina Nerghes. ۲۰۱۷. "Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N< ۱,۰۰۰)." *Journal of the Association for Information Science and Technology* ۶۸ (۴):۱۰۲۴-.۱۰۳۵

Li, Changzhou, Yao Lu, Junfeng Wu, Yongrui Zhang, Zhongzhou Xia, Tianchen Wang, Dantian Yu, Xurui Chen, Peidong Liu, and Junyu Guo. ۲۰۱۸. "LDA Meets Word۲Vec: A Novel Model for Academic Abstract Clustering." Companion of the The Web Conference ۲۰۱۸ on The Web Conference .۲۰۱۸

Li, Fang, and Xiaojie Wang. ۲۰۱۷. "Improving Word Embeddings for Low Frequency Words by Pseudo Contexts." In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, ۳۷-۴۷. Springer.

Liang, Jiye, Feng Wang, Chuangyin Dang, and Yuhua Qian. ۲۰۱۲. "An efficient rough feature selection algorithm with a multi-granulation view." *International Journal of Approximate Reasoning* ۵۳ (۶):۹۱۲–۹۲۶.

Liao, Wei-keng, Ying Liu, and Alok Choudhary. ۲۰۰۴. "A grid-based clustering algorithm using adaptive mesh refinement." ۷th Workshop on Mining

Scientific and Engineering Datasets of SIAM International Conference on Data Mining.

Lilleberg, Joseph, Yun Zhu, and Yanqing Zhang. ٢٠١٥. "Support vector machines and word٢vec for text classification with semantic features." Cognitive Informatics & Cognitive Computing (ICCI* CC), ٢٠١٥ IEEE ١٤th International Conference on.

Liu, Fang, Tor-Kristian Jenssen, Vegard Nygaard, John Sack, and Eivind Hovig. ٢٠٠٤. "FigSearch: a figure legend indexing and classification system." *Bioinformatics* ٢٠ (١٦):٢٨٨٠-.٢٨٨٢

Liu, Feifan, Deana Pennell, Fei Liu, and Yang Liu. ٢٠٠٩. "Unsupervised approaches for automatic keyword extraction using meeting transcripts." Proceedings of human language technologies: The ٢٠٠٩ annual conference of the North American chapter of the association for computational linguistics.

Liu, Huan, and Hiroshi Motoda. ١٩٩٨. *Feature extraction, construction and selection: A data mining perspective*: Springer.

Liu, Huan, and Hiroshi Motoda. ٢٠١٢. *Feature selection for knowledge discovery and data mining*. Vol. ٤٥٤: Springer Science & Business Media.

Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. ٢٠١٠. "Understanding of internal clustering validation measures." Data Mining (ICDM), ٢٠١٠ IEEE ١٠th International Conference on.

Liu, Yang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. ٢٠١٥. "Topical Word Embeddings." AAAI.

Luukka, Pasi, Kalle Saastamoinen, and Ville Könönen. ٢٠٠١. "A classifier based on the maximal fuzzy similarity in the generalized Lukasiewicz-structure." Fuzzy Systems, ٢٠٠١. The ١٠th IEEE International Conference on.

Lv, Peng Hui, Gui-Fang Wang, Yong Wan, Jia Liu, Qing Liu ,and Fei-cheng Ma. ٢٠١١. "Bibliometric trend analysis on global graphene research." *Scientometrics* ٨٨ (٢):٣٩٩-.٤١٩

Ma, Jing, and Alan L Porter.۲۰۱۵ "Analyzing patent topical information to identify technology pathways and potential opportunities." *Scientometrics*:۱-.۱۷

Ma, Long, and Yanqing Zhang. ۲۰۱۵. "Using Word۲Vec to process big text data." Big Data (Big Data), ۲۰۱۵ IEEE International Conference on.

Madnick, Stuart, Wei Lee Woon, Andreas Henschel, Ayse Firat, Blaine Ziegler, Steven Camina, Satwik Seshasai, Georgeta Vidican, Hatem Zeineldin, and Toufic Mezher. ۲۰۰۸. Technology Forecasting Using Data Mining and Semantics. MIT/MIST Collaborative Research.

Manomaisupat, P, and K Ahmad. ۲۰۰۵. "Feature Selection for Text Categorisation Using Self-Organising Map." Neural Networks and Brain, ۲۰۰۵. ICNN&B'۰۵. International Conference on.

Marcin, Ivan, and Sam Shiu. ۲۰۱۲. Extracting Topic Trends and Connections: Semantic Analysis and Topic Linking in Twitter and Wikipedia Datasets. Informally published manuscript, Computer Science, Stanford University, Retrieved from http://www. stanford. edu/class/cs۲۲۴w/upload/cs۲۲۴w-۰۰۴-final. v۰۱. pdf.

Martino, Joseph P. ۱۹۹۳. *Technological forecasting for decision making*: McGraw-Hill, Inc.

Martino, Joseph P. ۲۰۰۳. "A review of selected recent advances in technological forecasting." *Technological Forecasting and Social Change* ۷۰ (۸):۷۱۹-.۷۳۳

Mathioudakis, Michael, and Nick Koudas. ۲۰۱۰. "Twittermonitor: trend detection over the twitter stream." Proceedings of the ۲۰۱۰ ACM SIGMOD International Conference on Management of data.

Matsuo, Yutaka, and Mitsuru Ishizuka. ۲۰۰۴. "Keyword extraction from a single document using word co-occurrence statistical information." *International Journal on Artificial Intelligence Tools* ۱۳ (۰۱):۱۵۷–۱۶۹.

McDowall, William, and Malcolm Eames. ۲۰۰۶. "Forecasts, scenarios, visions, backcasts and roadmaps to the hydrogen economy: A review of the hydrogen futures literature." *Energy Policy* ۳۴ (۱۱):۱۲۳۶-.۱۲۵۰

Mihalcea, Rada, and Paul Tarau. ۲۰۰۴. "TextRank: Bringing order into texts".

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. ۲۰۱۳. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:*۱۳۰۱..۳۷۸۱

Mnih, Andriy, and Koray Kavukcuoglu. ۲۰۱۳. "Learning word embeddings efficiently with noise-contrastive estimation." Advances in neural information processing systems.

Mohamad, Ismail Bin, and Dauda Usman. ۲۰۱۳. "Standardization and its effects on K-means clustering algorithm." *Research Journal of Applied Sciences, Engineering and Technology* ۶ (۱۷):۳۲۹۹-.۳۳۰۳

Mohr, John W, and Petko Bogdanov. ۲۰۱۳. Introduction—Topic models: What they are and why they matter. Elsevier.

Müller, Andre Matthias, Carol A Maher, Corneel Vandelanotte, Melanie Hingle, Anouk Middelweerd, Michael L Lopez, Ann DeSmet, Camille E Short, Nicole Nathan, and Melinda J Hutchesson. ۲۰۱۸. "Physical Activity, Sedentary Behavior, and Diet-Related eHealth and mHealth Research: Bibliometric Analysis." *Journal of medical Internet research* ۲۰ (۴):e.۱۲۲

Nayak, Janmenjoy, Bighnaraj Naik, and HS Behera. ۲۰۱۵. "Fuzzy C-means (FCM) clustering algorithm: a decade review from ۲۰۰۰ to ۲۰۱۴." In *Computational intelligence in data mining-volume* ۲, ۱۳۳-۱۴۹. Springer.

Newman, Nils C, Alan L Porter, David Newman ,Cherie Courseault Trumbach, and Stephanie D Bolan. ۲۰۱۴. "Comparing methods to extract technical content for technological intelligence." *Journal of Engineering and Technology Management* ۳۲:۹۷-.۱۰۹

Nikitinsky, Nikita, Dmitry Ustalov, and Sergey Shashev" .۲۰۱۵ .An information retrieval system for technology analysis and forecasting." Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), .۲۰۱۵

No, Hyun Joung, and Yongtae Park. ۲۰۱۰. "Trajectory patterns of technology fusion: Trend analysis and taxonomical grouping in nanobiotechnology." *Technological Forecasting and Social Change* ۷۷ (۱):۶۳-.۷۵

Norton, Melanie. ۲۰۰۰. *Introductory concepts in information science*: Information Today, Inc.

Oelze, Iryna. ۲۰۰۹. Automatic Keyword Extraction for Database Search. Leibniz University. Hannover.

Ogburn, William Fielding. ۱۹۳۷. "Technological trends and national policy".

Ogura, Hiroshi, Hiromi Amano, and Masato Kondo. ۲۰۰۹. "Feature selection with a measure of deviations from Poisson in text categorization." *Expert Systems with Applications* ۳۶ (۳):۶۸۲۶-.۶۸۳۲

Ojo, Adebola K, and Adesesan B Adeyemo. ۲۰۱۵. "Trend Analysis in Academic Journals in Computer Science Using Text Mining." *International Journal of Computer Science and Information Security* ۱۳ (۴):.۸۴

Okada, Makoto, Kazuaki Ando, Samuel Sangkon Lee, Yoshitaka Hayashi, and Jun-ichi Aoe. ۲۰۰۱. "An efficient substring search method by using delayed keyword extraction." *Information processing & management* ۳۷ (۵):۷۴۱-.۷۶۱

Ozaydin, Bunyamin, Ferhat Zengul, Nurettin Oner, and Dursun Delen. ۲۰۱۷. "Text-mining analysis of mHealth research." *mHealth* .(۱۲)۳

Pal, Nikhil R, and James C Bezdek. ۱۹۹۵. "On cluster validity for the fuzzy c-means model." *IEEE Transactions on Fuzzy systems* ۳ (۳):۳۷۰-.۳۷۹

Peng, Yi, Yong Zhang, Gang Kou, Jun Li, and Yong Shi. ۲۰۱۲. "Multicriteria decision making approach for cluster validation." *Procedia Computer Science* ۹:۱۲۸۳-.۱۲۹۱

Peng, Yi, Yong Zhang, Gang Kou, and Yong Shi. ۲۰۱۲. "A multicriteria decision making approach for estimating the number of clusters in a data set." *PLoS one* ۷ (۷):e.۴۱۷۱۳

Phillips, Joanne G, Ted R Heidrick, and Ian J Potter. ۲۰۰۷. "Technology futures analysis methodologies for sustainable energy technologies." *International Journal of Innovation and Technology Management* ۴ (۰۲):۱۷۱-.۱۹۰

Pierre, John M. ۲۰۰۲. "Mining knowledge from text collections using automatically generated metadata." In *Practical Aspects of Knowledge Management*–۵۳۷ , .۵۴۸Springer.

Porter, Alan L. ۱۹۹۱. *Forecasting and management of technology*. Vol. ۱۸: John Wiley & Sons.

Porter, Alan L, and Scott W Cunningham. ۲۰۰۵. "Tech mining." *Competitive Intelligence Magazine* ۸ (۱):۳۰-.۳۶

Quan, Xiaojun, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. ۲۰۱۵. "Short and Sparse Text Topic Modeling via Self-Aggregation." IJCAI.

Rahayu, Endang Sri Rusmiyati, and Zainal A Hasibuan. ۲۰۰۶. "Identification of technology trend on Indonesian patent documents and research reports on chemistry and metallurgy fields".

Rajaraman, Kanagasabi, and Ah-Hwee Tan. ۲۰۰۱. "Topic detection, tracking, and trend analysis using self-organizing neural networks." Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Rendón, Eréndira, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. ۲۰۱۱. "Internal versus external cluster validation indexes." *International Journal of computers and communications* ۵ (۱):۲۷-.۳۴

Rokach, Lior. ۲۰۱۰. "Ensemble-based classifiers." *Artificial Intelligence Review* .۳۹–۱:(۲–۱) ۳۳

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. ۲۰۱۰. "Automatic keyword extraction from individual documents." *Text Mining*:۱-.۲۰

Rosenberg, Deborah. ۲۰۰۷. "Trend Analysis and Interpretation: Key Concepts and Methods for Maternal and Child Health Professionals".

Rotolo, Daniele, Diana Hicks, and Ben R Martin. ۲۰۱۵. "What is an emerging technology?" *Research Policy* ۴۴ (۱۰):۱۸۲۷-.۱۸۴۳

Rousseeuw, Peter J. ۱۹۸۷. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* ۲۰:۵۳-.۶۵

Roussel, Philip A. ۱۹۹۱. *Third generation R&D: managing the link to corporate strategy*: Harvard Business Press.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. " .۱۹۸۶Learning representations by back-propagating errors." *nature* ۳۲۳ (۶۰۸۸):.۵۳۳

Samadikuchaksaraei, Ali, Hafez Mohammadhassanzadeh, and Farhad Shokraneh. ۲۰۱۳. "A bibliometric trend analysis of stem cells and regenerative medicine research output in Iran :comparison with the global research output".

Schnaars, Steven P, Swee L Chia, and Cesar M Maloles III. ۱۹۹۳. "Five modern lessons from a ۵۵-year-old technological forecast." *Journal of Product Innovation Management* ۱۰ (۱):۶۶-.۷۴

Schubert, Erich, Michael Weiler, and Hans-Peter Kriegel. ۲۰۱۴. "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds." Proceedings of the ۲۰th ACM SIGKDD international conference on Knowledge discovery and data mining.

Schubert ,Erich, Michael Weiler, and Arthur Zimek. ۲۰۱۵. "Outlier detection and trend detection: two sides of the same coin." Data Mining Workshop (ICDMW), ۲۰۱۵ IEEE International Conference on.

Segaran, Toby. ۲۰۰۷. *Programming collective intelligence: building smart web* ۲.۰ *applications*: " O'Reilly Media, Inc.".

Shearer, Colin. ۲۰۰۰. "The CRISP-DM model: the new blueprint for data mining." *Journal of data warehousing* ۵ (۴):۱۳-.۲۲

Shen, Jun, Xiaoxia Li, and Zusha Gu. ۲۰۱۳. "Strategic Diagram Analysis Based on Knowledge Network." ۲۰۱۲ First National Conference for Engineering Sciences (FNCES ۲۰۱۲).

Smalheiser, NR. ۲۰۰۱. "Predicting emerging technologies with the aid of text-based data mining: the micro approach." *Technovation* ۲۱ (۱۰):۶۸۹-.۶۹۳

Small, Henry, Kevin W Boyack, and Richard Klavans. ۲۰۱۴. "Identifying emerging topics in science and technology." *Research Policy* ۴۳ (۸):۱۴۵۰-.۱۴۶۷

Sridhar, Vivek Kumar Rangarajan. ۲۰۱۵. "Unsupervised topic modeling for short texts using distributed representations of words ".Proceedings of the ۱st workshop on vector space modeling for natural language processing.

Steinbach, Michael, George Karypis, and Vipin Kumar. ۲۰۰۰. "A comparison of document clustering techniques." KDD workshop on text mining.

Strader, C Rockelle. " .۲۰۰۹Author-assigned keywords versus Library of Congress Subject Headings: Implications for the cataloging of electronic theses and dissertations." *Library resources & technical services* ۵۳ (۴):۲۴۳-.۲۵۱

Suh, Yongyoon, and Jeonghwan Jeon. ۲۰۱۸. "Monitoring patterns of open innovation using the patent-based brokerage analysis." *Technological Forecasting and Social Change*.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. ۲۰۱۴. "Learning sentiment-specific word embedding for twitter sentiment classification." Proceedings of the ۵۲nd Annual Meeting of the Association for Computational Linguistics (Volume ۱: Long Papers).

Tao, Hongzhi, Jianfeng Li, Tao Luo, and Cong Wang. ۲۰۱۷. "Research on topics trends based on weighted K-means." Electronics Information and Emergency Communication (ICEIEC), ۲۰۱۷ ۷th IEEE International Conference on.

Tseng, Yuen-Hsien, Chi-Jen Lin, and Yu-I Lin. ۲۰۰۷. "Text mining techniques for patent analysis." *Information Processing & Management* ۴۳ (۵):۱۲۱۶-.۱۲۴۷

Tu, Yi–Ning, and Jia-Lang Seng. ۲۰۱۲. "Indices of novelty for emerging topic detection." *Information processing & management* ۴۸ (۲):۳۰۳-.۳۲۵

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. ۲۰۱۰. "Word representations: a simple and general method for semi-supervised learning." Proceedings of the ۴۸th annual meeting of the association for computational linguistics.

Van Der Heijden, Kees. ۲۰۰۰. "Scenarios and forecasting: two perspectives." *Technological Forecasting and Social Change* ۶۵ (۱):۳۱-.۳۶

van Raan, Anthony. ۱۹۹۹. "Advanced bibliometric methods for the evaluation of universities." *Scientometrics* ۴۵ (۳):۴۱۷-.۴۲۳

Ververidis, Dimitrios, and Constantine Kotropoulos. ۲۰۰۸. "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition." *Signal Processing* ۸۸ (۱۲):۲۹۵۶-.۲۹۷۰

Vidican, Georgeta, Wei Lee Woon, and Stuart Madnick. ۲۰۰۹. "Measuring innovation using bibliometric techniques: The case of solar photovoltaic industry".

Viedma-Del-Jesus, Maria Isabel, Pandelis Perakakis, Miguel Ángel Muñoz, Antonio Gabriel López-Herrera, and Jaime Vila. ۲۰۱۱. "Sketching the first ۴۵ years of the journal Psychophysiology (۱۹۶۴–۲۰۰۸): A co-word-based analysis." *Psychophysiology* ۴۸ (۸):۱۰۲۹-.۱۰۳۶

Wang, Peng ,Jiaming Xu, Bo Xu, Chenglin Liu, Heng Zhang, Fangyuan Wang, and Hongwei Hao. ۲۰۱۵. "Semantic clustering and convolutional neural network for short text categorization." Proceedings of the ۵۳rd Annual Meeting of the Association for Computational Linguistics and the ۷th International Joint Conference on Natural Language Processing (Volume ۲: Short Papers).

Wang, Qi. ۲۰۱۸. "A bibliometric model for identifying emerging research topics." *Journal of the Association for Information Science and Technology* ۶۹ .۳۰۴–۲۹۰:(۲)

Wang, Yi, and Xiao-Jing Wang. ۲۰۰۵. "A New Approach to feature selection in Text Classification." Machine Learning and Cybernetics, ۲۰۰۵. Proceedings of ۲۰۰۵ International Conference on.

Wang, Zhibo, Long Ma, and Yanqing Zhang. ۲۰۱۶. "A hybrid document feature extraction method using latent Dirichlet allocation and word۲vec." Data Science in Cyberspace (DSC), IEEE International Conference on.

Wartena, Christian, Rogier Brussee, and Wout Slakhorst. ۲۰۱۰. "Keyword extraction using word co-occurrence." ۲۰۱۰ Workshops on Database and Expert Systems Applications.

Wei, Tingting, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. ۲۰۱۵. "A semantic approach for text clustering using WordNet and lexical chains." *Expert Systems with Applications* ۴۲ (۴):۲۲۶۴–۲۲۷۵.

Wilks, Daniel S. ۲۰۱۱. "Cluster analysis." In *International geophysics*, ۶۰۳-۶۱۶. Elsevier.

Woon, Wei Lee, Andreas Henschel, and Stuart Madnick. ۲۰۰۹. "A framework for technology forecasting and visualization." Innovations in Information Technology, ۲۰۰۹. IIT'۰۹. International Conference on.

Woon, Wei Lee, and Stuart Madnick. ۲۰۰۹. "Asymmetric information distances for automated taxonomy construction." *Knowledge and information systems* ۲۱ (۱):۹۱-.۱۱۱

Woon, Wei Lee, Hatem Zeineldin, and Stuart Madnick. ۲۰۱۱. "Bibliometric analysis of distributed generation." *Technological Forecasting and Social Change* ۷۸ (۳):۴۰۸-.۴۲۰

Wu, Feng-Shang, Chun-Chi Hsu, Pei-Chun Lee, and Hsin-Ning Su. ۲۰۱۱. "A systematic approach for integrated trend analysis—The case of etching." *Technological Forecasting and Social Change* ۷۸ (۳):۳۸۶-.۴۰۷

Wu, Jingqiao, Yanchun Liang, Xiaoyue Feng, and Gang Song. ۲۰۱۸. "Exploring Trends of Lung Cancer Research Based on Word Representation." International Conference on Artificial Intelligence: Methodology, Systems, and Applications.

Wu, Kuo-Lung. ۲۰۱۲. "Analysis of parameter selections for fuzzy c-means." *Pattern Recognition* ۴۵ (۱):۴۰۷-.۴۱۵

Xie, Xuanli Lisa, and Gerardo Beni. ۱۹۹۱. "A validity measure for fuzzy clustering." *IEEE Transactions on Pattern Analysis & Machine Intelligence* (۸):۸۴۱-.۸۴۷

Yan, Jun, Ning Liu, Benyu Zhang, Shuicheng Yan, Zheng Chen, Qiansheng Cheng, Weiguo Fan, and Wei-Ying Ma. ۲۰۰۵. "OCFS: optimal orthogonal centroid feature selection for text categorization." Proceedings of the ۲۸th annual international ACM SIGIR conference on Research and development in information retrieval.

Yau, Chyi-Kwei, Alan Porter, Nils Newman, and Arho Suominen. ۲۰۱۴. "Clustering scientific documents with topic modeling." *Scientometrics* ۱۰۰ (۳):۷۶۷-.۷۸۶

Yoon, Byungun, Rob Phaal, and David Probert. ۲۰۰۸. "Morphology analysis for technology roadmapping: application of text mining." *R&d Management* ۳۸ (۱):۵۱-.۶۸

Yoon, Janghyeok, and Kwangsoo Kim. ۲۰۱۲. "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection." *Scientometrics* ۹۰ (۲):۴۴۵-.۴۶۱

Zhang, Dongwen, Hua Xu, Zengcai Su, and Yunfeng Xu. ۲۰۱۵. "Chinese comments sentiment classification based on word۲vec and SVMperf." *Expert Systems with Applications* ۴۲ (۴):۱۸۵۷-.۱۸۶۳

Zhang, Yi, Jie Lu, Feng Liu, Qian Liu, Alan Porter, Hongshu Chen, and Guangquan Zhang. ۲۰۱۸. "Does deep learning help topic extraction? A kernel k-means clustering method with word embedding." *Journal of Informetrics* ۱۲ (۴):۱۰۹۹-.۱۱۱۷

Zhao, Chunhui, Xiaocui Li, and Yan Cang. ۲۰۱۵. "Bisecting k-means clustering based face recognition using block-based bag of words model." *Optik-International Journal for Light and Electron Optics* ۱۲۶ (۱۹):۱۷۶۱-.۱۷۶۶

Zhao, Fangkun ,Bei Shi, Ruixin Liu, Wenkai Zhou, Dong Shi, and Jinsong Zhang. ۲۰۱۸. "Theme trends and knowledge structure on choroidal neovascularization: a quantitative and co-word analysis." *BMC ophthalmology* ۱۸ (۱):۸۶

Ziegler, Blaine E. ۲۰۰۹. "Methods for bibliometric analysis of research: renewable energy case study." Massachusetts Institute of Technology.

سهیلی, فرامرز, علی اکبر خاصه, and پریوش کرانیان. ۱۳۹۷. "روند موضوعی مفاهیم حوزه علم اطلاعات و دانش‌شناسی ایران بر اساس تحلیل هم‌رخدادی واژگان." فصلنامه مطالعات ملی کتابداری و سازمان‌دهی اطلاعات ۲ (۲۹).

صدیقی, مهری. ۱۳۹۳. "بررسی کاربرد روش تحلیل همرخدادی واژگان در ترسیم ساختار حوزههای علمی)مطالعۀ موردی: حوزۀ اطلاعسنجی(" پردازش و مدیریت اطلاعات ۲ (۳۰):۲۵.

صدیقی, مهری, and عمار جلالی‌منش. ۱۳۹۱. "مطالعه روند پژوهش در حوزه مدیریت دانش در بازه زمانی ۲۰۰۱ تا ۲۰۱۰ و ترسیم ساختار آن." پردازش و مدیریت اطلاعات ۲ (۲۸):۲۰.