علوم اطلاع‌رسانی، دوره ۲۰، شماره ۱ و ۲ (پاییز و زمستان ۱۳۸۳) ۶۷-۹۵

آزمایش‌هایی دربارهٔ تأثیر تحلیل گفتمان بر الگوریتم‌های رده‌بندی و بازیابی اطلاعات[1]

جی. موراتو، جی. لورنس، ج..جنوا، جی. ا.موریرو

ترجمه محمدعلی ایمان‌پور

کارشناس ارشد ادیان و عرفان دانشگاه تهران

چکیده

پژوهشگران نظام‌های نمایه‌سازی و بازیابی، به‌منظور بهبود نتایج [جستجو]، همواره از گنجاندن اطلاعات بافتاری[2] بیشتر، پشتیبانی کرده‌اند. افزایش شمار پایگاه‌های اطلاعاتی متن کامل و پیشرفت‌های به دست آمده در ظرفیت ذخیره‌سازی رایانه‌ای، تحلیل متن را با بهره‌گیری از دانش زبان‌شناسی و فرا- زبان‌شناسی[3] امکان‌پذیر ساخته است. از میانهٔ دههٔ ۱۹۸۰، پژوهشگران توجه و گرایش بیشتری به بافتار پیدا کرده‌اند، و تحلیل گفتمان نقش مهم‌تری یافته است. هدف پژوهش توصیف شده در این مقاله، بررسی این مسئله است که آیا متغیرهای گفتمان، بر الگوریتم‌های نوین بازیابی و رده‌بندی اطلاعات اثر می‌گذارند یا نه. به‌منظور ارزیابی این فرضیه، چارچوبی عملی برای تحلیل اطلاعات در محیطی خودکار پیشنهاد شده است. در این محیط، ان- گرام‌ها[4] (فیلترکردن) و کی- مینز (میانگین عدد k) [5] و الگوریتم‌های رده‌بندی و چن با زیرمجموعه‌هایی از مدارک، برپایهٔ متغیرهای گفتمانی «گونه»[6]، «سیاق»[7]، «اصطلاح‌شناسی حوزه» و «ساختار مدرک» مورد آزمون قرار گرفتند. نتایج حاصل از مطالعهٔ الگوریتم‌های زیرمجموعه‌های مختلف، با ساختار اطلاعات « *سرعنوان‌های موضوعی پزشکی*» (مِش)[8] مقایسه شد. این نتایج نشان می‌دهد که ان- گرام‌ها وابستگی واضحی به متغیرهای گفتمان ندارند؛ هرچند که الگوریتم رده‌بندی کی- مینز چنین وابستگی را، البته فقط در «اصطلاح‌شناسی حوزه» و «ساختار مدرک» نشان می‌دهد، و سرانجام این‌که «الگوریتم چن» وابستگی مشخصی به همهٔ متغیرهای گفتمان دارد. از این اطلاعات می‌توان برای طراحی بهتر الگوریتم‌های رده‌بندی که باید متغیرهای گفتمان را مورد توجه قرار دهند، استفاده کرد. نتایج فرعی دیگری نیز از این پژوهش حاصل شده است که در مقاله ارائه می‌گردد.

کلیدواژه‌ها: الگوی گفتمان/ تحلیل بافتار/ زبان‌شناسی رایانه‌ای/ روش‌های تحلیل متن/ فیلترکردن/ ان –گرام‌ها/ کی – مینز/ هم- عبارت‌سازی[9]

مقدمه

از آن هنگام‌که «گارفیلد» (Grafield, 1953) در دههٔ پنجاه، اثر خود را پدید آورد، تحلیل زبان‌شناختی همواره با بهبود و اصلاح ابزارهای اطلاعاتی پیوند داشته است. امروزه، تحلیل ریخت‌شناسانه[10] نحوی و معناشناختی در نظام‌های تجاری بازیابی اطلاعاتِ نیز گنجانده می‌شوند (Warner,1994)، اما رهیافت‌های متن‌گرایانه هنوز پرشمار نیستند.

در آغاز دههٔ هفتاد، علاقهٔ زبان‌شناسان به بهبود تحلیل متن، اهمیت و ارزش متن را برجسته ساخت (Dijk, 1988). این گرایش امروزه تأثیر روزافزونی بر حوزه‌های دیگر، همچون نمایه‌سازی و فیلترکردن خودکار (Llorens, Velasco, Morato, & Moreiro, 1998)، یا ترجمهٔ ماشینی داشته است. در این رویکرد، فهم کامل یک جمله مستلزم آن است که برخی واژه‌ها، چون ضمایر و قیدها، در ارتباط با جمله‌های دیگر تفسیر شوند تا از این طریق موقعیت‌های مرجع‌دار[11] (کلماتی چون ضمیر و قید که به واژهٔ دیگری که پیشتر در گفتمان بیان شده اشاره دارند) روشن گردند. برخی پژوهش‌ها نشان می‌دهند که اشارات مرجع‌دار تأثیر مستقیمی بر عملکرد ابزارهای «پردازش زبان طبیعی» (ان‌ال‌پی)[12] به جا می‌گذارند. این موضوع را «میتکوف» (Mitkov, 1998) در بحث تحلیل ضمیر در ترجمهٔ ماشینی، یا «لورنس» و

همکاران (Liorens et.al., 1998) در بحث الگوریتم‌های رده‌بندی اطلاعات مورد بررسی قرار دادند.

با این‌که رویکرد تحلیل متن ناب، هنوز برای فهم این که متون از نظر اطلاعاتی، شیوه‌های بیان و سبک‌شناسی (Swales,1996) چگونه سازمان می‌یابند ضرورت دارد، تحلیل متن در رویکرد کل‌نگر به نظام‌های بازیابی اطلاعات، نابسنده است. ویژگی‌های خاصی چون اقدامات استنادی[۱۳] در مقاله‌های دانشگاهی را، فقط به مدد اطلاعات ورامتنی[۱۴] می‌توان تشخیص داد. همراه با نخستین آثاری که در دهۀ هفتاد در موضوع تحلیل خودکار گفتمان منتشر گردید، شیوه‌ای برای حل این معضل ارائه شد (Pêcheux, 1969). گفتمان در چارچوب قواعدی که رشته‌های علمی و گروه‌های اجتماعی آن‌ها را می‌آفرینند، عمل می‌کند. این گروه‌های اجتماعی در شماری از موارد واژگانی خاص، شکل‌های زبانی، قواعد نظام‌بخش و مفاهیم فرهنگی، وجه اشتراک دارند. در این بافتار بعضی از عناصر تحلیل گفتمان، نظیر ساختارهای متنی، گونه‌ها و سیاق‌ها به‌کار گرفته می‌شوند تا تحلیل متن، کامل گردد (Karlgren & Cutting, 1994). همۀ این عناصر در این مقاله بررسی می‌شوند تا اهمیت نسبی آن‌ها در ابزارهای بازیابی اطلاعات مطالعه گردد.

در بخش ۲، یک مرور نظری دربارۀ اصطلاح‌شناسی گفتمان ارائه خواهد شد. بخش‌های بعدی مقاله به شرح نوعی روش‌شناسی برای مطالعۀ تأثیر وجوه مختلف گفتمان بر نظام‌های بازیابی اطلاعات خواهد پرداخت.

۲. مرور نظری برخی از نمودهای گفتمان

در این مقاله، گفتمان را به مثابه نمونه‌ای از کاربرد زبان در نظر می‌گیریم که می‌توان نوع آن را بر پایۀ عواملی چون: گونه، سیاق، اصطلاح‌شناسی حوزه، یا ساختار مدرک طبقه‌بندی کرد. تحلیل گفتمان، سازمان زبان را در سطحی بالاتر از جمله یا پاراگراف مطالعه می‌کند، و از این‌رو توجه خود را به واحدهای زبانی بزرگتری معطوف می‌سازد.

تحلیل گفتمان، گسترۀ پهناوری است، ولی تاکنون بخش کوچکی از حوزۀ تعریف شدۀ زبان‌شناسی بوده است. یکی از دلایل این وضعیت، این است که مفهوم گفتمان بر رویکردهای مختلفی مبتنی است که از تعدادی از رشته‌های دانشگاهی گرفته شده‌اند (Schriffrin, 1994; Beghtol, 2001). اگرچه بسیاری از حرفه‌مندان زبان‌شناسی، اصطلاح‌شناسی «گفتمان» را دارای ابهام و پیچیدگی می‌دانند، با این حال برای ورود به تحلیل گفتمان دو نگرش به گونۀ کلی، پذیرفته شده است: نگرش ساختاری و نگرش کارکردی. نگرش ساختاری به متن می‌پردازد، قواعد ساخت آن را آشکار می‌سازد و واحدهای [زبانی] آن را تحلیل می‌کند. نگرش کارکردی به بافتار و سبک گفتمان می‌پردازد و زبان را در ارتباط با نقش اجتماعی آن تبیین می‌کند. در این مقاله، رویکردی کل‌نگر، که در بردارندۀ هر دو نگرش پیش‌گفته است، برای تحلیل گفتمان ارائه می‌شود.

مرور نظری برخی از جنبه‌های گفتمان[۱۵] در بخش‌های بعدی این مقاله ارائه شده‌اند. این جنبه‌ها دو نوع تحلیل ساختاری و کارکردی را توأماً دربردارند. در این بخش‌ها از چهار جنبۀ گفتمان بحث می‌شود: گونه، سیاق، اصطلاح‌شناسی حوزه، و ساختار مدرک. به‌نظر می‌رسد که همۀ این جنبه‌ها قویاً با یکدیگر ارتباط دارند.

۲- ۱. گونه

گونه [= نوع] را به‌طور کلی می‌توان چنین تعریف کرد: «تعدادی از رویدادهای ارتباطی که در مجموعه‌ای از مقاصد ارتباطی، اشتراک دارند» (Swales, 1990). اما همۀ دانشمندان با این تعریف، کاملاً موافق نیستند: بعضی از نویسندگان «گونه» را بخشی از مفهوم سیاق می‌دانند (Amitāy, 1998) (که بعداً شرح داده می‌شود). مقاصد ارتباطی توسط اعضای

جامعهٔ دانشگاهیان و حرفه‌مندانی که آن مقاصد دائماً در آن پدیدار می‌شوند، تعریف و به شکل متقابل ادارک می‌گردند. اقسام گونه‌ها عبارت‌اند از: برنامه‌های پخش خبر، دستورالعمل‌ها، مصاحبه‌های مطبوعاتی، احکام عمومی مذهبی، و مانند آن‌ها. پژوهشگران در چند سال اخیر، مطالعات بسیاری دربارهٔ گونهٔ اینترنت انجام داده‌اند. در آزمایش‌های دیگر، تحلیل موشکافانه‌ای همراه با شاخص‌های متعدد به‌کار گرفته شده‌اند تا اقسام گونه‌ها از هم بازشناسانده و متمایز شوند (Karlgren & Cutting, 1994; Morato, 1999).

در مبحث گونه، «مفهوم سنخ‌شناسی مدرک»[۱۶] معنای خاص‌تری دارد. نمونهٔ سنخ‌شناسی‌های مدرک عبارت‌اند از: یادداشت‌های پژوهش و پیشرفت کار[۱۷]، مقاله‌های پژوهشی، صورت‌جلسه‌های سخنرانی، و مانند این‌ها. پژوهشگران علم کتابداری و اطلاع‌رسانی به‌صورت گسترده‌ای از سنخ‌شناسی مدرک در کار خود بهره می‌گیرند. این پژوهشگران از مدت‌ها پیش، رده‌بندی خودکار این نوع سنخ‌شناسی‌ها را مورد مطالعه قرار داده‌اند. هدف آنان از این نوع پژوهش‌ها افزایش نسبت مانعیت- جامعیت و نسبیت ربط[۱۸] [در بازیابی اطلاعات] است. مثلاً «گیلیاروسکی»، «اوزیلووسکی» و «مودروف» (Gilyareusky et.al., 1997) برپایهٔ طول عنوان مقاله‌های نشریات کشاورزی، به بررسی رده‌بندی خودکار آن‌ها پرداخته‌اند. «هاس»، «سوگارمن» و «تیبو» (Hass et.al., 1996) نیز با بهره‌گیری از رده‌بندی خودکار واژگان شاخص در مقاله‌های تجربی، یک فیلتر متن[۱۹] ایجاد کردند. در این مقاله، میزان همانندی بین دو خوشه سنجیده می‌شود: خوشهٔ اول متشکل از آن دسته از واژگان تجربی است که از مدارک استخراج شده‌اند، و خوشهٔ دوم مشتمل بر واژگان مدارکی است که باید مورد آزمایش قرار می‌گرفته‌اند. در این تحقیق می‌خواهیم با بهره‌گیری از گونه، به سنخ‌شناسی مدرک بپردازیم.

۲-۲. سیاق

به‌منظور دادن تعریفی از سیاق، ضرورت دارد که بحث را با تعریف سبک آغاز کنیم. «کارلگرن» (Kavlgren, 1998) سبک را مجموعهٔ گزینه‌هایی می‌داند که از میان ساختارهای واژگانی، ساختارهای نحوی ـ صرفی، و نشانگرهای زبانیِ مختلف انتخاب می‌شوند تا به وسیلهٔ آن‌ها موضوع خاصی در مدارک بحث ارائه شود. بنابراین، سبک در هر زبان، از امکان انتخاب شکل‌های بدیل بیان[۲۰] سرچشمه می‌گیرد، شکل‌هایی که مشخص کنندهٔ یک شخص، گروه، یا دورهٔ زمانی خاصی هستند. تفاوت‌های سبکی[۲۱] به عوامل گوناگونی بستگی دارند، که اصلی‌ترین آن‌ها عبارت‌اند از: نخست، مخاطبان موردنظر و آن محیط گفتمان که محل تولید متن است، و دوم ترجیحات مؤلف و خصیصه‌های شخصی او. از همین‌رو، سبک ربط مستقیمی با فرضیهٔ «گافمن» دربارهٔ نقاط منحنی «زیف»[۲۲] (Egghe & Roussau, 1990) که از طریق واژه‌های کم‌بسامد مشخص می‌شوند، دارد.

پژوهشگران زبان‌شناسی اجتماعی معمولاً سبک را همان سیاق توصیف می‌کنند. سیاق یک وجه بافتاری است که میان گروهبندی‌های ویژگی‌های زبانی و ویژگی‌های موقعیتی متکّرر، همبستگی برقرار می‌سازد (Halliday, 1985). سیاق، گزینه‌های سبکی تعمیم یافته‌تری را بازنمایی می‌کند. همان‌گونه که «هالیدی» می‌گوید، زبان انسان بر سه نوع کارکرد اصلی مبتنی است: زمینه[۲۳]، فحوا[۲۴] و شیوهٔ گفتمان[۲۵].

اگرچه این‌کارکردهای سه‌گانه با سیاق، بستگی دارند، فحوا[ی زبان] یکی از مهم‌ترین آن‌ها است، زیرا از پیش، انتخاب گزینه‌های زبانی در مؤلفهٔ میان‌فردی را تعیین می‌کند (Lavid, 1995).

سیاق، کاملاً با دیگر نمودهای گفتمان همبسته است. مثلاً کاربردهای متفاوت زبان، که با نام زبان فرعی شناخته می‌شود، چه‌بسا از ماهیت مطلب، موضوع یا سیاق برآمده باشند. «لوسی» (Lossee, 1996) مشخصه‌های دستوری کلمات و عبارت‌های مرکّب را در

رشته‌های مختلف مطالعه کرده تا از این طریق، زبان فرعی مورد استفاده را مشخص سازد.

۲ـ۳. اصطلاح‌شناسی حوزه

در بافتار این مقاله، تعریف «حوزه»، دقیقاً با «زمینهٔ گفتمان» مرتبط است. خاستگاه «حوزه» را در قلمرو مهندسی نرم‌افزار می‌توان یافت و هدفش تعیین چارچوبی است که مجموعه‌ای از برنامه‌های کاربردی نرم‌افزاری در آن به اجرا درآید. مثلاً حوزهٔ بانکداری اشاره به بازنمایی همهٔ اطلاعاتی دارد که برای فهم و ایجاد برنامه‌های کاربردی بانکداری مورد نیازند. «اصطلاح‌شناسی حوزه»[۲۶] به مجموعه اصطلاحاتی دلالت دارد که حوزهٔ خاصی از دانش را به بهترین شکل توصیف می‌کنند. مقبولیت اصطلاح‌شناسی حوزه در هر قلمرو خاص، به سطح پختگی حوزه و همچنین به کوشش‌های انسانی برای گروهبندی اصطلاحات مختلف دریک واژگان، وابسته است. در آثار گوناگون، بویژه آثاری که به عرصهٔ مهندسی دانش/ نرم‌افزار تعلق دارند، چنین نتیجه گرفته شده که اصطلاح‌شناسی حوزه، تأثیر ژرفی بر عملکرد الگوریتم‌های رایانه‌مبنا دارد. «پریتو دیاز» (Prieto-Diaz, 1988) به‌کارگیری شیوه‌های تحلیل حوزه (دی‌ای)[۲۷] را فقط برای حوزه‌های پخته پیشنهاد می‌کند.

انتظار می‌رود که اصطلاح‌شناسی حوزه بر ساخت خودکار حوزه تأثیر عمیقی داشته باشد؛ این امر با شناسایی مفاهیمی (اصطلاحاتی) مربوط می‌گردد که حوزه و روابط میان آن‌ها را توصیف می‌کنند. این فناوری معمولاً به نام «دی‌ای» شناخته شده است. در زمینهٔ ساخت خودکار حوزه‌ها در چند سال گذشته آزمایش‌هایی انجام گرفته است (Neighbors, 1981; diáz, Velasco, Llorens, and Martinez, 1998). «لورنز»، «ولاسکو» و «مارتینز اورگا» (1997) این فناوری را برای ساخت خودکار اصطلاحنامه‌ها مورد استفاده قرار دادند. این روش با بهره‌گیری از مجموعه‌ای از شیوه‌های فیلترکردن متن، به تشخیص اصطلاحات معنادار می‌پردازد (Frakes & Baeza - Yates, 1992). اصطلاحات برگزیده معمولاً به بهترین شکل، مجموعه‌های مدارک را بازنمایی می‌کنند. سپس به‌وسیلهٔ الگوریتمی، اصطلاح‌ها به منظور ایجاد روابط سلسله‌مراتبی و افقی، خوشه‌بندی می‌شوند. علاوه براین پژوهش‌ها، «پولانکو»، «گریول» و «رویات» (Polanco, Grivel, & Royoute, 1995) نیز یک روش خوشه‌بندی را به منظور تشخیص متغیرهای کتاب‌سنجی به‌کار گرفتند تا اشکال گوناگون اصطلاحات را مورد بررسی تاریخی قرار دهند. «کالون»، «کورتیال»، و «پنان» (Callon, Courtial, & penan, 1993) در مطالعات خود به بررسی تحول گرایش‌های پژوهشی در زمینهٔ تحلیل خوشه‌ای پرداختند. نتایجی که «لوز» و «لماری» (Looze & Lemarie, 1997)دربارهٔ هم-عبارت‌بندی[۲۸] به‌دست آوردند نشان می‌دهد که برای دست‌یابی به تصویر کامل و جامع از یک حوزه، باید به تعداد نسبتاً زیادی از پایگاه‌داده‌ها استفاده کرد و باید مجموعه مطالب زیادی را تحلیل کرد.

۲ ـ ۴. ساختار مدرک

برای تخصیص معناشناسی کلی به یک مجموعهٔ کلمات، آن کلمات باید دارای ساختار بنیادین زبانی در تمام متن باشند (Leydesorff, 1997)، که معمولاً چنین ساختاری را ساختار مدرک[۲۹] می‌نامند. مدارکی که به‌گونه‌های خاصی (مثل مقاله‌های پژوهشی) تعلق دارند، غالباً بسیار ساخت‌یافته و مقید به قواعدی هستند. نویسندگان متعددی (Dijk, 1988; Hearst & Plaunt, 1993) پیشنهاد می‌کنند که کلیدواژه‌ها یا اصطلاحات توصیفی که نمایانگر نمایهٔ مدرک هستند به ساختار مدرکی که این کلمات و کلیدواژه‌ها در آن‌ها پدیدار شده‌اند، پیوند زده شوند. «کاندو» (Kando, 1997) نیز نشان داد که به‌کارگیری ساختارهای در سطح متن در جستجو، موجب افزایش میزان مانعیت در نظام‌های بازیابی اطلاعات می‌شود. فراتر از آن،

«لیدسدورف» (Leydesdorf, 1997) ادعا کرده که الگوهای هم‌غیابی[۳۰] و هم‌وقوعی[۳۱] واژه‌ها در هر بخش، شکل خاصی دارند.

آرایش دقیق اطلاعات در داخل ساختار متن، عامل ارزشمندی خواهد بود. آزمایش‌های شناختی نشان می‌دهند که راهبردهای رایجی که برای تشخیص اطلاعات ارزشمند به‌کار روند، به‌ندرت با مطالعهٔ متوالی مقاله‌ها حاصل می‌شوند. نتایج بعضی از مطالعات (Swales, 1990) گویای این مطلب‌اند که جستجوگران در وهله اول به دیدن چکیدهٔ مقاله‌ها و سپس نتیجه‌گیری، گرایش دارند و پس از آن به سراغ تصاویر و جدول‌ها، و سرانجام به سراغ یافته‌ها می‌روند.

۳. چارچوب آزمایشی برای بررسی تأثیر متغیرهای گفتمان در الگوریتم‌های نمایه‌سازی و رده‌بندی

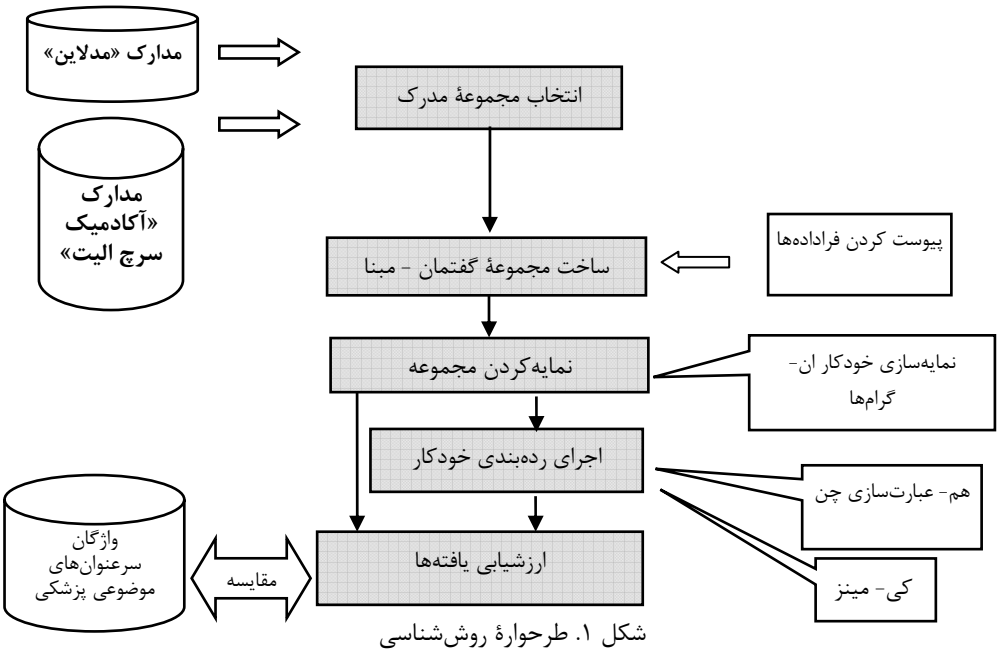
در این بخش، روش‌ها و چارچوب آزمایشی برای ارزیابی تأثیر متغیرهای گفتمان در محدودهٔ فنون دانش اطلاع‌رسانی ارائه می‌گردند. هدف اصلی، تدارک شیوه‌ای است تا تأثیر گونه‌ها، سیاق‌ها، اصطلاح‌شناسی‌های حوزه یا بخش‌های مختلف مدرک بر دو نوع الگوریتم متفاوت، یعنی بر نمایه‌سازی و رده‌بندی، ارزیابی شوند. مرحلهٔ نخست این روش‌شناسی (شکل ۱) ایجاد مجموعه‌ای از مدارک الکترونیکی است که از پایگاه داده‌ای مدارک گردآوری شده‌اند. برای یکپارچه‌سازی ساختار همهٔ مدارک گردآمده، باید این مدارک، پیشاپیش پردازش شوند. این مجموعه بر اساس متغیرهای مختلف گفتمان به زیرمجموعه‌هایی تقسیم می‌گردند. در مراحل بعدی برای مقایسهٔ یافته‌های مربوط به انواع گونه‌ها، سیاق‌ها، اصطلاح‌شناسی‌های حوزه و ساختارهای مدرک، الگوریتم‌های نمایه‌سازی و رده‌بندی در مورد همهٔ زیرمجموعه‌های پیش‌گفته استفاده می‌شوند. نتایج حاصل از کاربست این الگوریتم‌ها در زیرمجموعه‌های گفتمان- مبنا، باید با ساختار کاملاً تعریف‌شدهٔ اطلاعات (مانند اصطلاحنامهٔ سرعنوان‌های موضوعی پزشکی (مِش)) مقایسه شود تا تأثیر متغیرهای گوناگون گفتمان در الگوریتم، مورد ارزیابی قرار گیرد. معیار ارزیابی چنین بود: «نتایج این الگوریتم‌ها هرقدر به ساختار نزدیکتر باشند، الگوریتم‌ها بهتر اجرا شده‌اند.» همان‌گونه که در بخش چهار بحث شده است، دلیل کاربست چنین معیاری این بود که روابط [معنایی اصطلاحات] در اصطلاحنامهٔ «مش»، در میان متخصصان پزشکی به گونهٔ گسترده پذیرفته شده است. از سوی دیگر، تصوّر براین بود که خوب است با الگوریتم‌های کلاسیک بازیابی اطلاعات، مقایسه‌هایی صورت گیرد تا از این رهگذر ارزش آن‌ها در کنار اصطلاحنامهٔ «مش» نیز سنجیده شود (به‌طور مثال به بخش «۴-۱» نگاه کنید). یافته‌های قطعی (این که کدام متغیر بهتر از همه عمل می‌کند) و بویژه یافته‌های نسبی (مقایسهٔ یافته‌های حاصل از ارزش‌های مختلف متغیرها در داخل زیرمجموعه‌ها) به ما این امکان را می‌دهد که تأثیر متغیرهای گوناگون گفتمان را بر الگوریتم‌های مورد مطالعه اندازه‌گیری کنیم. این روش‌شناسی در بخش‌های بعدی نشان داده خواهد شد.

۳ـ ۱. انتخاب مجموعهٔ مدارک

تعداد برگزیدهٔ مدارک، از ۴۴۰ مدرک الکترونیکی متن کامل تشکیل می‌شد. این مدارک از دو پایگاه متفاوت داده‌ها استخراج شدند: «مدلاین» [نظام پیوسته بازکاوی مدارک پزشکی] و «آکادمیک سرچ الیت»[۳۲]. این پایگاه‌ها به علت دسترسی‌پذیری و داشتن مقاله‌های متن کامل از ناشران معتبر، انتخاب شدند.

این مدارک برای آزمودن متغیرهای زیر برگزیده شدند: ۱) پنج گونهٔ مختلف («مقاله‌های پژوهشی»، «اخبار»، «صورت‌جلسه‌های کنفرانس»، «یادداشت‌ها»، و «مقاله‌های علمی عامه‌پسند»[۳۳]، ۲) سه سیاق مختلف («زبان علمی»، «زبان مطبوعاتی» و «زبان علمی عامه‌پسند»، ۳) پنج حوزهٔ گوناگون اصطلاح‌شناسی («بیماری هپاتیت»، «اچ‌آی‌وی» [= ویروس ایدز]، «سی‌جی‌دی»، «پروتئین‌های گیاهی»، و «پروتئین‌های بالینی» که خود از دو حوزهٔ جداگانه (پزشکی و

زیست‌شناسی) برگرفته شده‌اند زیرا این دو حوزه، سازمان و واژگان بسیار قاعده‌مندی دارند (Nwogo, 1997)، و ۴) شش بخش متفاوت مدارک («چکیده»، «مقدمه»، «روش‌ها»، «یافته‌ها»، «بحث»، و «ارجاعات»).



شکل ۱. طرحوارهٔ روش‌شناسی

آثار چاپی زیر، از سال ۱۹۹۶ و از پایگاه داده‌های «مدلاین» انتخاب شدند:

- New England Journal of Medicine
- British Medical Journal, Lancet
- Journal of Clinical Investigation
- AIDS Care

از پایگاه «آکادمیک سرچ»، انتشارات زیر برمبنای گونه، بازیابی شدند:

- US News & World Report
- The Economist
- Newsweek
- Time

به عنوان مقاله‌های پژوهشی:

- Plant Physiology

به عنوان مقاله‌های علمی عامه‌پسند:

- Blood Weekly

۳ـ ۲. ساخت مجموعهٔ فرعی گفتمان - مبنا

برای تحلیل تأثیر متغیرهای مختلف گفتمان بر الگوریتم‌ها، یک معیار فرعی‌تر برای مدارک منتخب (یکی از چهار متغیر گفتمان مذکور در زیر) به‌کار گرفته شد: گونه، سیاق، اصطلاح‌شناسی حوزه، یا ساختار مدرک. متناظر با هر یک از متغیرهای گفتمان، چهار گروه مختلف از مدارک، ایجاد گردید. هر گروه دارای ۴۲۴ مدرک مشابه بود (غیر از گروه ساختار مدرک که از ۴۲۴ مدرک پیش‌گفته، فقط ۹۴ مدرک داشت)، که متناسب با ارزش متغیر گفتمان، در زیرمجموعه‌هایی گروه‌بندی شدند. مثلاً در گروه «گونه»، ۴۲۴ مدرک به پنج دسته تقسیم گشتند؛ در این حالت تمام مقاله‌های پژوهشی با هم در یک زیرمجموعه قرار گرفتند، همهٔ اخبار در زیرمجموعه دیگر، و بقیهٔ مدارک نیز به همین نحو تقسیم‌بندی شدند. مدارک به روش دستی به هر مجموعه اختصاص یافتند. نتیجهٔ کار در نهایت، نوزده زیرمجموعهٔ زیر بود که هر یک دارای مدارک همگنی بودند:

گروه «گونه» (پنج زیرمجموعه): «مقاله‌های پژوهشی»، «اخبار»، «صورت‌جلسه‌های کنفرانس»، «یاداشت‌ها»، «مقاله‌های علمی عامه‌پسند».

گروه «سیاق» (سه زیرمجموعه): «زبان علمی»، «زبان مطبوعاتی» و «زبان علمی عامه‌پسند».

گروه «اصطلاح‌شناسی‌های حوزه» (پنج زیرمجموعه): «هپاتیت»، «اچ‌آی‌وی»، «سی‌جی‌دی»، «پروتئین‌های گیاهی»، «پروتئین‌های بالینی».

گروه «ساختارهای مدرک» (شش زیرمجموعه): «چکیده»، «مقدمه»، «روش‌شناسی»، «یافته‌ها»، «بحث»، «ارجاعات».

۳ـ۲ـ۱. زیرمجموعه‌های «گونه»

پنج زیرمجموعهٔ «گونه» بر پایهٔ سنخ‌شناسی مدرک و نشریه‌ای که مدارک از آن گرفته شده بودند، ساخته شد. «یادداشت‌ها» و «صورت‌جلسه‌های کنفرانس» از «مدلاین» اخذ شدند، یعنی از همان جایی‌که «اخبار» و «مقاله‌های علمی عامه‌پسند» انتخاب شده بودند. «مقاله‌های پژوهشیِ» «آکادمیک سرچ الیت» هم در «*مدلاین*» یافت شدند و هم در «آکادمیک سرچ الیت». مقاله‌های پژوهشی باید مشخصه‌های علمی می‌داشتند و توسط داوران مستقل، ارزیابی می‌شدند. مقاله‌های کوتاه، از ویژگی‌های علمی برخوردار نبودند اما در انتشارات علمی با عنوان «یادداشت‌ها» ذکر شده بودند. مدارکی که محتوای بسیار تازه‌ای داشتند ولی مفاهیم مذکور در مقاله را توصیف نمی‌کردند، به گونهٔ «اخبار» اختصاص می‌یافتند. شماری از مدارک روزنامه‌ای («یو اس نیوز» و مانند آن) اگر دارای معیارهای زیر بودند، به روش دستی به «مقاله‌های علمی عامه‌پسند» اضافه می‌شدند: ۱) هدفشان توضیح یک موضوع علمی مشخص برپایه مبانی اولیه، برای خواننده غیرمتخصص باشد؛ ۲) مقاله، اطلاعات حاصل از نتیجهٔ یک تحقیق بلندمدت را که انتظار می‌رود تأثیر آن بر حوزهٔ دانش، ادامه‌دار باشد، توصیف می‌کند. از این نظر، سرمقاله‌ها مورد توجه نبودند.

ساختار این زیرمجموعه در جدول شمارهٔ ۱ ارائه شده است.

۳ـ۲ـ۲. زیرمجموعه‌های سیاق

به منظور ایجاد زیرمجموعه‌های سیاق، همهٔ مدارک خوانده شدند و به شیوهٔ ذهنی، به متغیرهای مختلف سیاق منسوب شدند. این یافته‌ها به دست آمد: همهٔ مقاله‌های پژوهشی در زیرمجموعهٔ «زبان علمی» جا داده شدند. تقریباً همهٔ «صورت‌جلسه‌های کنفرانس‌ها» و «یادداشت‌ها» به «زبان علمی» نسبت داده شدند، و بقیه در دستهٔ «علم عامه‌پسند» جای گرفتند. بخشی از مطالب خبری در گروه «علم عامه‌پسند» جای یافتند و بقیه به «زبان مطبوعاتی» منسوب شدند. بنابراین، میان گونه و سیاق، همخوانی دقیقی نتیجه نمی‌شد.

مدارکی که از روزنامه‌ها اخذ شده و هدف اصلی در آن‌ها توضیح اصطلاح‌شناسی علمی بود، در زمرهٔ «زبان علمی عامه‌پسند» قرار گرفتند. مقاله‌های نشریهٔ «بلاد ویکلی»[۳۴] نیز در زمرهٔ همان «زبان علمی عامه‌پسند» جای گرفتند، زیرا دارای همان ویژگی‌هایی بودند که «پوسترگیلو» (Posterguillo, 1996) مشخص کرده بود (یعنی پرکردن صفحه، و...).

ساختار این زیرمجموعه در جدول ۲ ارائه شده است.

جدول ۱. زیرمجموعه‌های گونه

نام نشریه	زیرمجموعه‌ها					
	سرمقاله	مقاله‌های علمی- عمومی	یادداشت‌ها	صورت جلسه‌های کنفرانس‌ها	اخبار	مقاله‌های پژوهشی
New England Journal of	۶		۸			۲۳
British Medical Journal	۱		۳۵			۲۶
Journal of Clinical Investigation						۴۸
Lancet	۷		۴۶			۳۵
AIDS Care			۵	۱		۴۳
US News & World Report		۱			۱۵	
The Economist	۱	۱			۱۱	
Newsweek	۱	۱	۱		۱۱	
Time		۲			۶	
Plant Physiology						۴۴
Blood Weekly		۳۶		۹		

جدول ۲. زیرمجموعه های سیاق

نام نشریه	زیرمجموعه‌ها		
	زبان علمی	زبان مطبوعاتی	زبان علمی عامه‌پسند
New England Journal of Medicine	۳۷		
British Medical Journal	۶۲		
Journal of Clinical Investigation	۴۸		
Lancet	۸۸		
AIDS Care	۴۹		
US News & World Report		۱۶	
The Economist	۴	۹	
Newsweek	۲	۱۲	
Time	۱	۷	
Plant Physiology	۴۴		
Blood Weekly			۴۵

۳ـ۲ـ۳. زیرمجموعه‌های «اصطلاح‌شناسی حوزه»

تقریباً همهٔ انتشارات اخذشده از «مدلاین» به‌گروه مضمونی «پزشکی عمومی و داخلی» «جی‌سی‌آر» تعلق دارد. مجلهٔ "Journal of Clinical Investigation" که به گروه مضمونی «آزمایش‌های پژوهشی- پزشکی» تعلق دارد، به‌این منظور انتخاب شد که تفاوت اصطلاح‌شناسی حوزه را در این مجموعه نشان دهد.

جدول ۳. زیرمجموعه‌های اصطلاح‌شناسی حوزه

نام نشریه	زیرمجموعه‌ها				
	هپاتیت	اچ‌ای‌وی	سی‌جی‌دی	پروتئین‌های گیاهی	پروتئین‌های
New England Journal of Medicine	۲	۲۴			
British Medical Journal	۱۳	۴۰	۹		
Journal of Clinical Investigation		۱			۴۷
Lancet	۲۲	۶۵	۱		
AIDS Care		۴۹			
US News & World Report		۱۶			
The Economist		۱۲	۱		
Newsweek		۱۴			
Time		۸			
Plant Physiology				۴۴	
Blood Weekly		۲۸			

در حوزهٔ پزشکی، معیار انتخاب مدرک، از سه بیماری عالم‌گیر، یعنی هپاتیت، ایدز، و بیماری کروتسفلد- یاکوب (سی‌جی‌دی)[۳۵] تشکیل می‌شد که برای مقایسهٔ مراحل مختلف تحقیق انتخاب شدند. هپاتیت بیماری بسیار قدیمی و دارای پراکندگی ناهمگنی است که در متون پزشکی، واژگان آن کاملاً قاعده‌مند شده‌اند. ایدز یک بیماری جهانی جدید و فراگیر در سال ۱۹۹۶ (زمان آغاز طراحی این کار) بود و سی‌جی‌دی نیز یک بیماری پراکنده و نادر بود که در سال ۱۹۹۶ در مناطق روستایی ظاهر شد و واژگانش در مراحل اولیه قرار داشت.

در حوزهٔ زیست‌شناسی، از رشته‌های زیست‌شیمی و گیاه‌شناسی ۹۱ مدرک بازیابی شد. اصطلاح‌شناسی‌های حوزه‌ای مختلفی که انتخاب شدند عبارت بودند از: پروتئین‌های گیاهی از "Plant Physiology Journal" و پروتئین‌های بالینی از "Journal of Clinical Investigation".

ساختار این زیرمجموعه در جدول ۳ ارائه شده است.

۴ـ۲ـ۳. زیرمجموعه‌های ساختار مدرک

به‌منظور مطالعهٔ این که آیا الگوریتم‌های نمایه‌سازی و رده‌بندی با شیوه‌ای متفاوت و بسته به آن بخشی از مدرک که با آن کار می‌کنند عمل می‌کنند یا نه، از مجموع ۴۲۴ مدرک، ۹۴ مدرک برپایهٔ سازمان آی‌ام‌آردی (مقدمه، روش، یافته‌ها، و بحث)،[۳۶] که توسط «بروس» پیشنهاد گردیده بود ساختاربندی شدند (Bruce, 1983). ساختاربندی شدند ۹۴ مقالهٔ علمی طبق شکل بسط‌یافته‌ای از ساختار آی‌ام‌آردی، که شامل چکیده و ارجاعات نیز بود، سازماندهی گردید. هرگاه مقاله‌های منتخب فاقد یکی از بخش‌های ساختار آی‌ام‌آردی[۳۷] بودند، جای آن بخش خالی گذاشته می‌شد. بقیهٔ مدارک بویژه آن‌هایی که متعلق به گونهٔ «اخبار» بودند، به خاطر اشکالاتی که در ساختاربندی آن‌ها با استفاده از ساختار بسط‌یافتهٔ آی‌ام‌آردی روی می‌داد، مورد توجه واقع نشدند. فرایند تخصیص ساختار به‌طور خودکار و با به‌کارگیری یک برنامهٔ انطباق رایانه- مبنا انجام شد. در این الگوریتم، اصطلاحات تکواژه‌ای موجود در عنوان بخش‌های مختلف مدرک، با سیاهه‌ای از اصطلاحات مربوط به هر بخش بسط‌یافتهٔ آی‌ام‌آردی[۳۸] انطباق داده می‌شد. بالاترین میزان انطباق در بخش‌های چکیده، مقدمه و ارجاعات به‌دست آمد. اما به‌منظور جلوگیری از اشتباهات، فرآیند انطباق را

نویسندگان این مقاله کاملاً کنترل می‌کردند. با این‌که از سازمان آی‌ام‌آردی به‌گونه‌ای گسترده در پزشکی و زیست‌شناسی استفاده می‌شود، اما درصد بالایی از ۹۴ مدرک، اساساً از این ساختار تبعیت نکرده بودند. برای تبدیل بخش‌های مختلف مدارک به سازمان آی‌ام‌آردی، نویسندگان از تجربه‌های پیشین «سوالز» (Swales, 1990)، «نووگو (Nwogo, 1997)، و «اسکلتون» (Sketton, 1990) استفاده کردند.

جدول ۴. زیرمجموعه‌های ساختار مدرک

نام نشریه	زیرمجموعه‌ها					
	چکیده	مقدمه	روش‌ها	یافته‌ها	بحث	ارجاعات
New England Journal of Medicine	۲۰	۲۰		۱۴	۱۷	۲۰
British Medical Journal	۲۲	۲۲	۲۲	۲۲	۲۲	۲۱
Lancet	۲۸	۲۸	۲۸	۲۸	۲۸	۲۸
AIDS Care	۲۴	۲۴	۲۴	۲۴	۲۱	۲۰

حاصل پردازش اولیه در این مرحله، گزیده‌ای از مدارک بود که با ساختار بسط‌یافتهٔ آی‌ام‌آردی تهیه شده بودند (جدول ۴).

۳ـ۳. آزمون الگوریتم‌های نمایه‌سازی‌ـ بازیابی

در روش‌های کلاسیک بازیابی اطلاعات، یکی از مهم‌ترین کارها مربوط به انتخاب اصطلاحات باربطی است که مدارک هدف را توصیف کنند. بر پایه این دیدگاه، هرگاه که باید با یک مدرک الکترونیکی کار کرد برنامه‌های رایانه‌ای می‌کوشند به شکل خودکار اصطلاحاتی را انتخاب کنند که باید از متن کامل، به عنوان «توصیفگرها»ی متن برگزیده شوند. این فرآیند معمولاً از طریق الگوریتم‌های ان‌ـگرام‌ها (Cohen, 1995)، حذف واژه‌های بازدارنده، الگوریتم‌های بسامد اصطلاح‌ـ بسامد معکوس مدرک (تی‌اف‌ـ آی‌دی‌اف)[۳۹] (Spark Jones, 1972)، یا صرفاً با حذف دستی به اجرا درمی‌آید و معمولاً فیلترکردن متن (Frakes & Baeza, 1990) نامیده می‌شود. ان‌ـ گرام‌ها یا تی‌اف‌ـ آی‌دی‌اف، الگوریتم‌هایی هستند که نه تنها اطلاعات بسامد مربوط به هر مدرک، بلکه اطلاعات بسامد بِین‌مدرکی[۴۰] را نیز به‌کار می‌گیرند. در این پژوهش، برای آزمون تأثیر متغیرهای گفتمان، الگوریتم ان‌ـ گرام‌ها انتخاب شده است.

۳ـ ۳ـ ۱. الگوریتم فیلترکردن اصطلاحِ ان‌ـگرام‌ها

الگوریتم ان‌ـگرام‌ها از طریق شکستن متن ورودی به گرام‌ها (یا گروه‌های) n نویسه‌ای، فیلترکردن اصطلاح را انجام می‌دهد. مثلاً: با به‌کارگیری یک الگوریتم پنج گرامی، متن ورودی «carbon monoxide» باید پنج گرام زیر به دست آید: «carbo»، «arbon»، «rbon »، «bon m»، « on mo»،...

این گرام‌ها از نظر آماری با مجموعهٔ گرام‌های یک مجموعهٔ پس‌زمینه‌ای مقایسه می‌شوند. فرآیند فیلترکردن با پذیرش اصطلاحاتی که در بردارندهٔ گرام‌های مورد قبول‌اند، ساخته می‌شود. بالاترین امتیاز این الگوریتم به مثابه یک نمایه‌ساز، این است که می‌تواند نه‌فقط تک‌واژه‌ها، بلکه عبارت‌های اسمی را به‌عنوان توصیفگر (فیلترکنندهٔ متن) انتخاب کند. اما بنا به تصمیماتی که دربارهٔ اندازهٔ گرام و نیز اندازه، رشته و جزئیات پس‌زمینه گرفته می‌شود، یافته‌های متفاوتی به دست می‌آیند (Díaz et al., 1998).

برای محاسبه و سنجش ربط هر گرام، از فرمول زیر استفاده می‌شود:

$$Y_i = \begin{cases} C_i \ln (C_i / S) + B_i \ln (B_i / R) - (SC_i + RB_i) \\ \ln [(SC_i + RB_i)(S+R)], \; SC_i > RB_i \;\; 0, \\ SC_i < RB_i \end{cases}$$

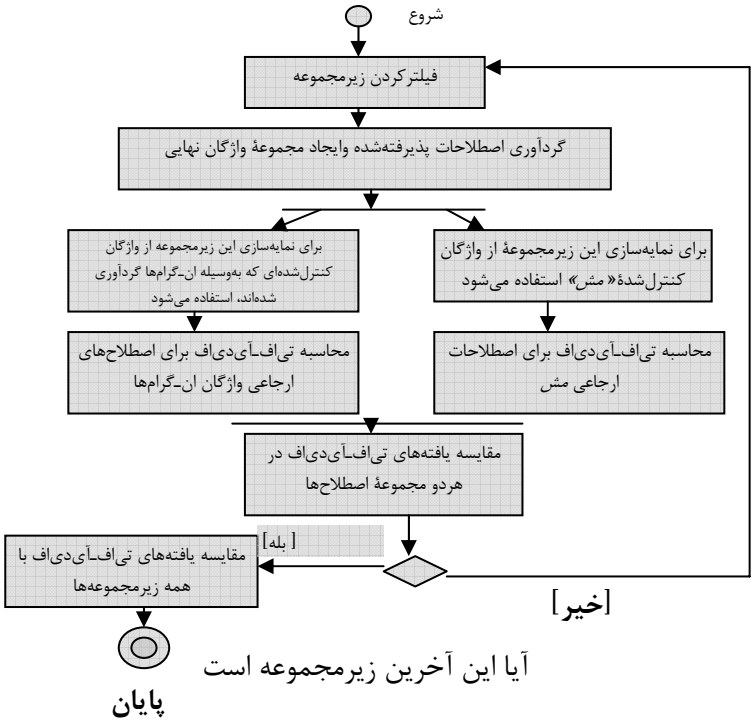
در این فرمول C_i نمایانگر ارزش i ان‌ـگرام در مدرک است، B_i نشانگر ارزش i ان‌ـگرام در پس زمینه، S نمایانگر ارزش مجموعهٔ ان‌ـگرام‌ها در این مدرک، و R نشانگر ارزش مجموعهٔ ان‌ـگرام‌ها در پس زمینه است. تصمیم‌گیری دربارهٔ طرح‌های مختلف باید در هنگام به‌کارگیری ان‌ـگرام‌ها مورد توجه قرار گیرد (Llorens et al., 1998). از جملهٔ این تصمیم‌ها این است که این مجموعه چگونه پردازش شود. ان‌ـگرام‌ها میان توصیفگرها تمایز قائل می‌شوند و فقط آن توصیفگرهایی را می‌پذیرند که دارای بسامدهای وقوع

متوسط[۴۱] هستند. همانطور که در «توزیع زیف» پیش‌بینی می‌شود، این بسامدهای متوسط، مفاهیم مدرک را بهتر از توصیفگرهای پربسامدتر یا کم‌بسامدتر بازنمایی می‌کنند. با این همه، واژه‌هایی که فقط یک‌بار در مدرک ظاهر می‌شوند ممکن است اطلاعات مهمی را بازنمایی کنند. تصمیم دیگری که اتخاذ شد، اندازهٔ ان‌ـگرام‌ها بود. آزمایش‌های ما نشان داده که ارزش‌های فرد سه تا هفت نویسه‌ای برای هر گرام، توصیفگرهای بهتری را حاصل می‌کنند.

همان‌گونه‌که ان‌ـگرام‌ها می‌توانند به مانند یک الگوریتم فیلترکردن اصطلاح عمل کنند، می‌توان از آن‌ها برای ایجاد مجموعه اصطلاح‌های نامزد که باید یک حوزه از پیکرهٔ الکترونیکی مدارک را بازنمایی کنند، بهره گرفت. این ویژگی با اندازه‌گیری کیفیت پیکرهٔ [اصطلاح‌های نامزد] در مقایسه با «*مش*» برای مقایسهٔ تأثیر متغیرهای مختلف گفتمان بر الگوریتم، به‌کار گرفته می‌شود.

۳ـ۳ـ۲. چارچوب آزمون ان‌ـگرام‌ها

ساختار این آزمایش برای سنجش الگوریتم ان‌ـگرام‌ها، در شکل ۲ ارائه شده است.



شکل ۲. ساختار آزمایشی برای سنجش الگوریتم ان ـ گرام‌ها

در مورد هر زیرمجموعه:

- باید با استفاده از ان – گرام‌ها، فیلترکردن انجام گیرد.
- از اصطلاحاتی که به وسیلهٔ ان‌ـگرام‌ها انتخاب شده‌اند، یک واژگان اصطلاحی ایجاد می‌شود.
- دو فعالیت به موازات هم انجام می‌شوند:

الف) نمایه‌سازی و محاسبهٔ تی‌اف‌ـ آی‌دی‌اف برای واژگانی که از طریق ان‌ـگرام‌ها گردآوری شده‌اند.

ب) نمایه‌سازی و محاسبهٔ تی‌اف‌ـآی‌دی‌اف برای واژگان «مش».

یافته‌های درون- زیر- مجموعه‌ای را (به منظور بررسی عملکرد کامل) با هم مقایسه کنید،

بعد از اقدام دربارهٔ همهٔ زیرمجموعه‌ها،

یافته‌های بین-زیر-مجموعه‌ای را (به منظور اندازه‌گیری تأثیر متغیرهای گفتمان بر این الگوریتم) با یکدیگر بسنجید.

وقتی الگوریتم فیلترکردن، اصطلاحات باربط مجموعه را انتخاب کرد، مجموعهٔ مدارک مورد ارجاع قرار گرفت. دو ارجاع نمایه‌ای ایجاد شد: اولی از واژگان ان-گرام‌ها به عنوان واژگان کنترل‌شده[۴۲] و دیگری از واژگان «مش» استفاده می‌کرد. هدف اصلی این فرآیند نیز تعیین کیفیت الگوریتم‌های ان-گرام در متغیرهای مختلف گفتمان بود. برای مقایسهٔ هر دو فرایند نمایه‌سازی پیش‌گفته، آزمون تی‌دی-آی‌دی‌اف به‌کار گرفته شد. اصل بنیادین آزمون آی‌دی‌اف این است که اهمیت یک اصطلاح در درون مدرک، اگر بسامدش در همهٔ مدارک در سطح پایینی باشد، بیش‌تر است. در بازیابی اطلاعات، ارزش بالای تی‌اف- آی‌دی‌اف برای یک اصطلاح، دلالت براین نکته دارد که باید آن اصطلاح را برای ساختن نمایهٔ آن مدرک انتخاب کرد. در فرآیند نمایه‌سازی، هر توصیفگر همراه با اطلاعاتی دربارهٔ تعداد دفعات وقوع آن و بخشی از مقاله که واژه در آن ظاهر شده، در پایگاه داده‌ها فهرست شد. توصیف مفصل‌تر این فرآیند را می‌توانید در اثر «دیاز» و دیگران (Diaz et al., 1998) بیابید.

۳-۴. آزمون الگوریتم‌های رده‌بندی

یکی از فعالیت‌های پژوهشی اصلی در علم اطلاع‌رسانی از نظر سازماندهی اطلاعات، پرداختن به ایجاد الگوریتم‌هایی است که به‌صورت خودکار، روابط میان اصطلاحات متن را درمی‌یابند. این الگوریتم‌ها را معمولاً الگوریتم‌های رده‌بندی می‌نامند. به منظور آزمون آن‌ها، میان روابطی که به‌وسیلهٔ این الگوریتم‌ها ایجاد می‌شوند با ساختار پذیرفته شدهٔ سلسله‌مراتبی از اصطلاحات (مانند «مش»)، مقایسه‌ای به عمل آمد.

دو الگوریتم معروف برای دستیابی به روابط میان اصطلاحات انتخاب شدند: الگوریتم «کی- مینز» و «الگوریتم چن». این دو روش بدین دلیل انتخاب شدند که نتایج خوبی در آثار پیشین به بار آورده بودند (Díaz et al., 1998).

۳-۴-۱. الگوریتم رده‌بندی کی- مینز

«کی- مینز» یکی از رایج‌ترین روش‌های خوشه‌بندی به‌شمار می‌آید. این الگوریتم معمولاً برای ایجاد خوشه‌هایی از اشیا که دارای ویژگی‌های مشترک هستند، به‌کار رفته است. این اشیا در علم اطلاع‌رسانی عبارت‌اند از: مدارک، کاربران، ارجاعات، جویش[۴۳]ها، یا همچون در پژوهش ما – خوشه‌هایی از اصطلاحات که در زیرمجموعه‌ها یافت می‌شوند.

این الگوریتم معمولاً برای تحقق رهیافت‌های بالا به پایین، که شناسایی سلسله‌مراتب‌ها را برای بازنمایی حوزه آسان می سازند، به‌کار می‌رود. روش کی- مینز به خانوادهٔ الگوریتم‌های تحلیل خوشه‌ای«مرکز سیار»[۴۴] تعلق دارد (Lelu, 1993). این به آن معنا است که مرکزیت هر گروه از اصطلاحات، پس از ورود یک مجموعهٔ جدید اصطلاحات مدرک، از نو طراحی می‌شود. کی- مینز متضمن تعدادی شاخص ورودی مهم است که برای کنترل فرآیند رده‌بندی به کار می‌روند، مثل تعداد مطلوب خوشه‌ها یا معیارهای تثبیت‌شده برای انتخاب توصیفگرهایی که شکل‌دهندهٔ ریشهٔ سلسله‌مراتب هستند.

۳-۴-۲. چارچوب آزمون کی- مینز

در چارچوبی که ما در نظر گرفته‌ایم، روش کی- مینز باید پس از فرآیند نمایه‌سازی محاسبه شود. ساخت سلسله‌مراتب‌ها با بهره‌گیری از رهیافت بالا به پایین انجام می‌شود. ابتدا باید یک ریشه انتخاب شود؛ پس از گزینش ریشه، باید با استفاده از بقیهٔ توصیفگرهایی که شیوهٔ کی- مینز را به کار گرفته‌اند، فرآیند خوشه‌بندی به اجرا درآید. درونداد روش کی- مینز، مجموعه بردارهایی است که هر یک از این بردارها برای یک اصطلاح از واژگانی درنظر گرفته می‌شود که براساس فرآیند فیلترکردن ان-گرام‌ها گرد آمده‌اند. هر مؤلفه از بردارها، تعداد نسبی وقوع اصطلاح در یک

مدرک بخصوص را نمایش می‌دهد. کاربرد کی- مینز برای مجموعهٔ بردارها، خوشه‌های مختلف، و نیز اطلاعات مربوط به مرکزیت هر خوشه را ایجاد می‌کند.

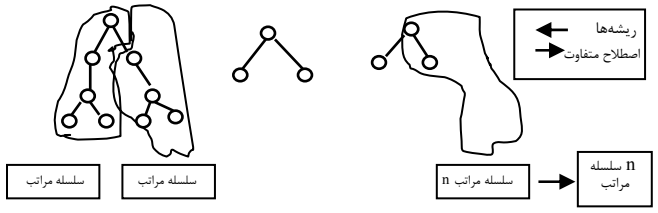
هنگامی‌که این فرایند پایان می‌یابد، با خوشه‌های مختلف همچون اصطلاحات خاصِ ریشهٔ قبلی رفتار شد. با استخراج اجزای سازندهٔ اصلی هر خوشه، دستیابی به سطح بعدی سلسله‌مراتب ممکن گردید. این ریشه‌های جدید، اصطلاحات خاصی بودند که پیش از ریشه‌های سطح اول قرار می‌گرفتند. روش‌هایی که برای استخراج ریشه‌ها به‌کار رفت عبارت بودند از: «فاصله از مرکزیت»، «شمار بیشینهٔ دفعات وقوع»، «شمار بیشینهٔ مدارک»، و «ضریب عمومیت» (Díaz et al., 1998).

به منظور بررسی تأثیر گفتمان بر کارآیی [الگوریتم] کی- مینز، یافته‌های این خوشه که به وسیلهٔ کی- مینز گرد آمده بودند، باید با سلسله‌مراتب‌های درختی «مش» مقایسه می‌شدند. یک مشکل عمده این است که کی- مینز به یک شاخص ورودی نیاز دارد، که به تعداد خوشه‌های مطلوب اشاره دارد. برآورد این مقدار، تصمیمی بسیار مهم است. اگر مقدار پایین باشد، الگوریتم تعداد اندکی خوشه خواهد ساخت و از همین‌رو احتمال این که هر خوشه دربردارندهٔ سلسله‌مراتب‌های «مش» باشد بسیار بالا خواهد بود. از سوی دیگر، مقدار بالای خوشه‌ها تلویحاً به این معنا است که تعداد کم‌تری از سلسله‌مراتب‌های «مش» در آن هستند. این مقدار را می‌توان با تعداد خوشه‌های یافت‌شده در «مش» برابر کرد. بنابراین، آزمایش باید ابتدا سلسله‌مراتب «مش» را تعیین کند (شکل ۲).

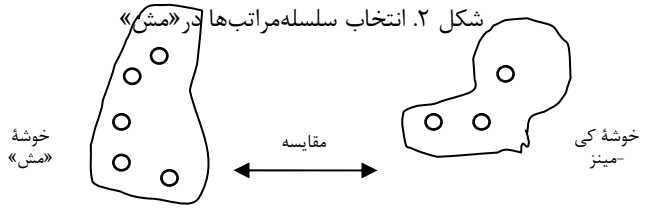
علاوه براین، کی- مینز به‌طور مستقیم سلسله‌مراتب‌ها را فراهم نمی‌سازد، بلکه اصطلاحات خوشه‌بندی‌شده، همهٔ اصطلاحات موجود در هر سلسله‌مراتب «مش»، در یک خوشه گروه‌بندی می‌گردند و سپس با خوشه‌های کی- مینز مقایسه می‌شوند (شکل ۳).

۳-۴-۳. الگوریتم رده‌بندی هم- عبارت‌سازی چن

الگوریتم چن نوعی از الگوریتم هم- عبارت‌سازی[۴۵] است. الگوریتم هم- عبارت‌سازی نتیجهٔ پیشنهادهای مختلفی است که با کتاب‌سنجی ارتباط دارند. این الگوریتم اساساً به دنبال ساخت نقشه‌های علم از طریق استخراج روابط میان وقوع واژه‌ها است (Callon, Coutial, Turner, and Bauin, 1983). فرضیهٔ اساسی در هم- عبارت‌سازی این است که اگر دو اصطلاح، معمولاً به‌صورت مشترک در یک مدرک واحد حضور داشته باشند از نظر معناشناسی با هم مرتبط هستند. بنابراین سنجش فاصلهٔ معنایی میان دو اصطلاح از طریق محاسبهٔ هم- وقوعی و هم- غیابی آن‌ها در یک مجموعهٔ مدارک، امکان‌پذیر است.



شکل ۲. انتخاب سلسله‌مراتب‌ها در «مش»



شکل ۳. چارچوب برای آزمون کی- مینز

روش چن با الگوریتم هم‌ـ عبارت‌سازی کار می‌کند. این الگوریتم برای هر جفت اصطلاح، یک ضریب ایجاد می‌کند که درجهٔ رابطهٔ [معنایی] (معمولاً همایندی) آن‌ها را نسبت به هم اندازه می‌گیرد. نتیجهٔ این الگوریتم، ماتریسی از روابط میان اصطلاحات است.

ضریبی برای هر جفت اصطلاح تعیین می‌شود که به وزن خوشه اشاره دارد.

عامل وزن (T_k) وزن خوشه $(Tj, Tk)=\frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}}$

عامل وزن (T_j) وزن خوشه $(Tk, Tj)=\frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{jk}}$

در فرمول بالا d_{ij} و d_{ik} همان تی‌اف‌ـ آی‌دی‌اف هستند و عبارت، همان مقدار تی‌اف‌ـآی دی اف را نشان می‌دهد، در حالی که دو اصطلاح j و k در مدرک واحد i هستند (هم‌ـ عبارت‌سازی)، و بنابراین اگر t_{fijk} تعداد وقوع هر دو اصطلاح j و k در مدرک I باشد، d_{fjk} تعداد مدارک دارای اصطلاح j و k را نشان می‌دهد.

برای هر جفت اصطلاح، ضریبی تعیین می‌شود تا وزن خوشه (عددی در حد فاصل ۱ و ۱-) بین این دو اصطلاح را نشان دهد. این ضریب با مقدار آستانه مقایسه می‌شود. بر مبنای تعداد آستانه، نوع رابطهٔ میان توصیفگرها استنتاج می‌گردد. با استفاده از مجموعهٔ توصیفگرها و مقادیری که از میزان آستانه [مقادیر حد بالا و حد پایین] درمی‌گذرند، می‌توان نموداری تهیه کرد.

روش چِن فقط به تعیین روابط میان اصطلاح‌ها گرایش دارد، اما میان انواع رابطه (هم‌ارزی، همایندی پایدار) تمایزی را مشخص نمی‌سازد. این روش فقط برای تعیین نوع کلی روابط، یعنی همایندی کارآمد است.

۳ـ۴ـ۴. چارچوب آزمون هم‌ـ عبارت‌سازی چن

در این طرح پژوهشی، «الگوریتم چن» پس از اجرای فرایند نمایه‌سازی به‌کار گرفته شد. برای هر جفت اصطلاح، ضریبی ساخته می‌شود که به سطح همایندی اصطلاحات اشاره دارد. این وزن با سنجش بسامد معکوس مدرک آی‌دی‌اف و بسامد درون مدرکی تی‌اف محاسبه می‌شود (Chen & Lynch, 1992).

ساختار این آزمون به شرح زیر بود:

مجموعهٔ اصطلاح‌هایی که در ان‌ـگرام‌ها ایجاد شده بود، به‌مثابه درونداد الگوریتم چن به‌کار رفت. نتیجه، گروهی از همایندی‌های دودویی میان اصطلاحات بود. این ارتباط‌ها با توجه به وابستگی‌های موجود در «مش» مورد بررسی قرار گرفتند.

۴. آزمایش‌ها و یافته‌ها

به منظور بررسی تأثیر متغیرهای گفتمان بر الگوریتم‌های بازیابی اطلاعات، سه آزمایش اصلی به اجرا در آمد:

آزمایش اول:

وابستگی الگوریتم ان‌ـگرام‌ها به متغیرهای گفتمان را ارزیابی کنید. در زیرمجموعه‌های گفتمان- مبنا، با هدف ساخت مجموعه‌ای از اصطلاحات که می‌توانند واژگان حوزه‌ای مربوط به هر زیرمجموعه را شکل دهند، الگوریتم فیلترکردن ان‌ـ گرام‌ها را به اجرا درآورید. سپس، با استفاده از اصطلاحات ان‌ـگرام‌ها به‌عنوان واژگان کنترل‌شده و بعد با استفاده از واژگان مش زیرمجموعه‌های مدرک را نمایه‌سازی کنید. ارزش‌های بسامد اصطلاح‌ـ بسامد معکوس مدرک را در مورد همهٔ اصطلاح‌های ارجاع داده شده محاسبه کنید، و یافته‌های بسامد مربوط به هر دو مجموعه واژگان کنترل‌شده را در مورد هر یک از زیرمجموعه‌های متغیر گفتمان به مقایسه بگذارید.

اهداف: ۱. کیفیت ان‌ـگرام‌ها را به‌عنوان الگوریتم فیلترکردن ایجاد واژگان، بسنجید.

۲. تأثیرمتغیرهای گفتمان بر الگوریتم ان ـگرام‌ها را ارزیابی کنید.

۳. وابستگی اصطلاح‌شناسی حوزه در دیگر متغیرهای گفتمان را بررسی کنید؛ بویژه چگونگی استفادهٔ حوزه‌های نزدیک به هم از مترادف‌های مختلف برای توصیف یک مفهوم واحد را مورد بررسی قرار دهید.

آزمایش دوم:

وابستگی الگوریتم کی‌ـ مینز به متغیرهای گفتمان را ارزیابی کنید. الگوریتم رده‌بندی کی‌ـ مینز را در مورد اصطلاح‌های گردآوری شده در زیرمجموعه‌های گفتمان- مبنا اجرا کنید، تا خوشه‌های اصطلاح- مبنا[۴۶] را ایجاد نمایید و یافته‌ها را با سلسله‌مراتب درختی «مش» مقایسه کنید.

اهداف: ۱. کیفیت کی‌ـ مینز را به‌عنوان الگوریتم رده‌بندی سلسله‌مراتب‌گرا[۴۷] اندازه‌گیری کنید.

۲. تأثیر متغیرهای گفتمان را بر الگوریتم کی‌ـ مینز ارزیابی کنید.

آزمایش سوم:

وابستگی الگوریتم چن به متغیرهای گفتمان را بسنجید. الگوریتم ارتباط چن را در زیرمجموعه‌های مختلف گفتمان- مبنا به اجرا بگذارید تا بدین ترتیب میان اصطلاحات، بستگی به وجود آید و یافته‌ها را با درختوارهٔ «مش» مقایسه کنید.

اهداف: ۱. کیفیت الگوریتم چن را به‌مثابه یک الگوریتم رده‌بندی بستگی‌ـ گرا[۴۸] بسنجید.

۲. تأثیر متغیرهای گفتمان در الگوریتم چن را ارزیابی کنید.

نویسندگان این مقاله پژوهشی دریافتند که یکی از جالب‌ترین ویژگی‌های چارچوب پیشنهادشده آن است که از الگوریتم‌های انتخاب شده، بسیار مستقل است. از آنجا که هدف اصلی این چارچوب، مقایسهٔ عملکرد الگوریتم‌ها در مواقعی است که در مجموعه‌های گفتمان‌ـ مبنا از مدارک به کار گرفته می‌شوند، جالب‌ترین نتایج را می‌توان از مقایسه‌های نسبی یافته‌ها به‌دست آورد.

برای این که فرآیندهای نمایه‌سازی و رده‌بندی تقویت شوند، اطلاعات خاصی در پایگاه داده‌هایی که باید توسط نمایه‌سازان و الگوریتم‌های رده‌بندی به‌کار گرفته شود، ذخیره شد:

- با استفاده از «پروژه اسمارت» سیاههٔ بازدارنده‌ای تهیه شد (SMART, 2000). پیش از نمایه‌سازی، مطابق این سیاهه، همهٔ واژه‌های بازدارنده حذف شدند.
- یک واژه‌نامهٔ انگلیسی به پایگاه داده‌ها افزوده شد. به هر مدخلی، برچسبی حاوی مقولهٔ دستوری آن تعلق گرفت. منبع این واژه‌نامه، «وردنت»[۴۹] و «مجموعهٔ ملی بریتانیایی»[۵۰] بود (BNC, 2001).
- سرعنوان‌های موضوعی معروف «مش» به این پایگاه داده‌ها ضمیمه شد. واژگان «مش» هر موضوعی را در پزشکی زیر پوشش می‌گیرد و ارزش آن در نمایه‌سازی، از سوی جامعه پزشکی به‌گونه‌ای گسترده به تأیید رسیده است. دو نسخهٔ پیراسته از این منبع، عملاً به این پایگاه داده‌ها افزوده گردید: «ساختار درختی مش»[۵۱] با سازمان سلسله‌مراتبی، و «فهرست الفبایی مشروح مش»[۵۲]، به همراه اصطلاحات مرتبط، مترادف‌ها، و املاهای مختلف هر توصیفگر. «واژگان مش» دربردارندهٔ ۱۸،۰۰۰ توصیفگر و ۱۰۰،۰۰۰ مترادف است (Lowe & Barnett, 1994). این واژگان معتبر، برای انجام دو کار بنیادین در این طرح پژوهشی به‌کار گرفته شد: اول این که سیاههٔ اصطلاحات «مش» با سیاههٔ اصطلاحاتی که از طریق فرآیند فیلترکردن همهٔ زیرمجموعه‌ها گرد آمده بود، مقایسه گردید، و سپس «ساختار درختی مش» با ساختارهای اصطلاحاتی که به وسیلهٔ الگوریتم‌های رده‌بندی ایجاد شده بودند، به سنجش گذاشته شد. اگر بتوان تنوع یافته‌ها را در زیرمجموعه‌های مختلف نشان داد، آنگاه می‌توان فرض کرد که ارتباط نمودهای گفتمان با

کاربرد فیلترکردن و الگوریتم‌های رده‌بندی، مطمئناً امکان‌پذیر است.

- قواعد ریشه‌یابی (که در بخش‌های بعدی این مقاله توصیف خواهند شد).

۴ـ۱. آزمایش اول: تأثیر گفتمان بر الگوریتم ان‌ـگرام‌ها

پس‌زمینه‌ای که در این مطالعه برگزیده شد از مقاله‌های زمین‌شناسی و یک رمان تاریخی تشکیل می‌شد. این گونه‌ها به این دلیل انتخاب شدند که با مجموعهٔ مدارک موضوع پژوهش، از نظر گفتمان همپوشانی اندکی داشتند. برای بهبود نتایج نمایه‌سازی، یک برنامهٔ رایانه‌ای، مدرک را بررسی می‌کرد تا نویسه‌هایی [= حروفی] را که در سامانهٔ ما دارای ارزش معنایی اندکی بودند حذف کند:

- نویسه‌های بی‌معنا، مانند زبرنوشت‌ها[۵۳]، اعداد پانوشت‌ها و فصول، نشانهٔ «@»، نشانه‌های عاطفی، و از این قبیل؛
- علائم نشانه‌گذاری (پرانتزها، علائم نقل‌قول، کج‌خط‌ها، ستاره، و مانند این‌ها).

برای این‌که همهٔ اصطلاحات فیلترشدهٔ ان‌ـگرام‌ها به شکل قاعده‌مند درهم ادغام شوند، یک الگوریتم ریشه‌یابی کلمات (Díaz, Llorens & Morato, 2002) طراحی و در سامانهٔ پژوهش اجرا شد. این سامانه انجامهٔ هر واژه را می‌یافت، آن را با فهرست انجامه‌های اصطلاحات موجود در پایگاه داده‌ها مقابله می‌کرد، و شکل قاعده‌مندی را جایگزین آن می‌ساخت: با این فرایند، همهٔ واژه‌هایی که هم‌ستاک [= هم‌ریشه] بودند در یک واژهٔ واحد ادغام می‌شدند. این اطلاعات می بایست توسط ریشه‌یاب‌ها، در دو جدول پایگاه داده‌ها گنجانده می‌شد: یعنی جدول «وندها»، و جدول «انجامه‌های جانشین». مثلاً کلمه «virus» و «viruses» مفهوم همسانی را بازنمایی می‌کردند و می‌توانستیم با انتخاب اصطلاح «virus» آن را قاعده‌مند کنیم.

اصطلاحاتی که برپایهٔ فرآیند بالا به‌دست آمدند، با توصیفگرهای مجموعهٔ واژگان کنترل‌شده مقایسه شدند. هر یک از این اصطلاحات که در واژگان کنترل‌شده یافت شد، توصیفگر و بسامد وقوع آن، در پایگاه‌داده‌ها روزآمد گردید. اگر اصطلاحی نه توصیفگر بود و نه در سیاههٔ بازدارنده قرار داشت، در فهرست «اصطلاحات نامزد» گذارده می‌شد. به‌منظور انجام موفقیت‌آمیز این فرآیند، افعال شناسایی شدند و با مقایسه‌ای که میان شکل قاعده‌مندشدهٔ آن‌ها با واژه‌نامه‌های «بی‌ان‌سی» و «وردنت» به عمل آمد، کنار گذاشته شدند.

در این بررسی، ارزش هر گرام نهایتاً ۵ درنظر گرفته شده بود.

وقتی همهٔ زیرمجموعه‌ها فیلتر شدند و اصطلاحات به‌دست‌آمده قاعده‌مند گردیدند، به‌وسیلهٔ رایانه تعداد ۱۷۴۸ اصطلاح انتخاب شد تا واژگان کنترل‌شدهٔ ان‌ـگرام‌ها را تشکیل دهند.

با ارجاع همهٔ اصطلاحاتی از هر مدرک که در واژگان‌های کنترل‌شده وجود داشت، فرآیند نمایه‌سازی زیرمجموعه‌ها انجام شد. تعداد دفعات وقوع هر اصطلاح و نیز موقعیت‌های مختلف آن‌ها در مدارک نیز، در یک پایگاه داده‌ها ذخیره شدند. نتایج نشان داد که نمایه‌سازی این مجموعه با به‌کارگیری واژگان ان‌ـگرام‌ها، ارجاعاتی را برای ۱۷۴۸ اصطلاح ایجاد کرد، در حالی که نمایه‌سازی که با به‌کارگیری واژگان «*مش*» انجام شده بود، ارجاعاتی (شامل ارجاعات مترادف) را برای ۳۶۲۵ اصطلاح تولید کرد.

بسامد اصطلاح ـ بسامد معکوس مدرک با محاسبهٔ اطلاعات مربوط به دفعات وقوع هر زیرمجموعه، برای همهٔ اصطلاحات نمایه‌سازی‌شده (۱۷۴۸ و ۳۶۵۲ واژه) محاسبه گردید. فرمول مورد استفاده به قرار زیر بود:

$$\mathrm{idf} = \log_2\left(\frac{N}{n_i}+1\right)$$

در این فرمول N مساوی با تعداد کل مدارک موجود در زیرمجموعه و n_i برابر با تعداد مدارک موجود در زیرمجموعه‌ای است که شامل اصطلاح i می‌باشد. این مقدار در بسامد اصطلاح (تی‌اف) ضرب شد تا مقدار نهایی به‌دست آید. همه قبول دارند که هر چه «بسامد اصطلاح- بسامد معکوس مدرک» بالاتر باشد، اصطلاح مورد نظر اطلاعات مدرک را بهتر نشان می‌دهد. بنابراین در مقایسهٔ «تی‌اف-آی‌دی‌اف» برای زیرمجموعه‌های مشابه، که اولی با به‌کارگیری واژگان ان-گرام‌ها و دیگری با به‌کارگیری «مش» انجام می‌شود، می‌توان میزان مشخصی از کیفیت را برای ان-گرام‌ها سنجید. اما با مقایسهٔ یافته‌های مربوط به زیرمجموعه‌های مختلف، می‌توان استنتاج کرد که چگونه متغیرهای گفتمان براین الگوریتم اثر می‌گذارند. به‌منظور دست‌یابی به مقدار واحد «تی‌اف– آی‌دی اف» در کل یک زیرمجموعه، میانگین «تی‌اف-آی‌دی اف» همهٔ اصطلاحات، محاسبه شد. این اندازه‌گیری، اطلاعات زیادی دربارهٔ اصطلاحات به ما نمی‌دهد، اما کمک می‌کند که رفتار زیرمجموعه‌ها و واژگان‌های مختلف را مقایسه کنیم.

نتایج به‌دست‌آمده در جدول ۵ نشان داده می‌شود.

توضیحات زیر را دربارهٔ این جدول می‌توان ارائه کرد:

۱. ستون آخر این جدول (تفاوت ان-گرام‌ها در «مش») نشان می‌دهد که الگوریتم ان-گرام‌ها در بررسی متغیرهای چهارگانهٔ گفتمان، نتایج بهتری به دست آوردند. این نکته بسیار جالب است، زیرا می‌توانیم نتیجه بگیریم که به جای استفاده از واژگان کنترل‌شدهٔ کاملاً معروف و تثبیت شده برای نمایه کردن مدارک هر حوزهٔ خاص، بهتر است برای دست‌یابی به اهداف و مقاصد بازیابی، واژگان کنترل شده به طور خودکار و متناسب با مجموعهٔ مورد نظر، ساخته شود، و آنگاه با استفاده از این واژگان به نمایه‌سازی مجموعه اقدام گردد. نویسندگان این مقاله توضیح کاملی برای این نتایج ندارند، اگر چه می‌توان آن را با نظام وزن‌دهی[۵۴] تلویحی که الگوریتم ان-گرام‌ها در هنگام ساخت واژگان به کار می‌بندند، مرتبط ساخت. الگوریتم ان-گرام‌ها در هر مجموعه، اصطلاحات «بی‌ربط» را حذف کرده و از همین‌رو، کاربرد «تی‌اف-آی‌دی‌اف» در مورد این گونه اصطلاحات، ممکن است نتایج بهتری به بار آورد.

با نگاهی به این نتایج، در مورد متغیرهای گفتمان درمی‌یابیم که بالاترین تفاوت در متغیر «سیاق» (بالاترین میانگین در ستون سوم رخ داده است)، که درعین‌حال دارای بهترین واریانس نیز هست.

در مورد متغیر «گونه» و به‌طور کلی در گونه‌هایی که جنبهٔ فنی کمتری دارند و به همین دلیل بیشتر در حاشیهٔ حوزهٔ پزشکی قرار می‌گیرند (مانند «مطبوعات» و «علم عامه‌پسند») این تفاوت‌ها، از نظر قدر مطلق، در بالاترین حد خود بود. هیچ تبیین روشن و دقیقی برای این نتایج به‌دست نیامده است.

۲. محاسبات میانگین آماری برای هر متغیر گفتمان: بالاترین مقادیر «تی‌اف-آی‌دی‌اف»که به‌وسیلهٔ اصطلاحات واژگان ان-گرام‌ها به دست آمد در «مقاله‌های مطبوعات» و «یادداشت‌ها»ی مربوط به گونه، در «زبان مطبوعاتی» سیاق، در «ایدز» و «پروتئین‌های بالینی» اصطلاح‌شناسی حوزه، و در «ارجاعات» و «چکیده» ساختار مدرک یافت شد. اما «مش» در «صورت‌جلسهٔ کنفرانس» و «علم عامه‌پسند» گونه، در «زبان علمی عامه‌پسند» سیاق، در «سی‌جی‌دی» اصطلاح‌شناسی حوزه، و در «ارجاعات» و «چکیده» ساختار مدرک، نتایج بهتری را کسب کرد.

باید به یاد داشته باشیم که ان-گرام‌ها با استفاده از بسامد گرام‌ها، گرام‌های پذیرفته شده را برمی‌گزینند. این ویژگی می‌تواند مبین این نکته باشد که چرا این الگوریتم، مقادیر بالایی را در متغیرهای مختلف گفتمان به‌دست آورده است. بالاترین مقدار در «ایدز» به‌دست آمد، و به سبب انعکاس این بیماری در مطبوعات و انتشارات سال ۱۹۹۶ (همان سال انتشار مدارک)، این

وضعیت قابل توجیه بود. تأثیر اصطلاحاتی چون «ایدز» و «اچ‌آی‌وی» در مطبوعات بسیار بالا بود.

جدول ۵. مقایسه «تی‌اف-ای‌دی‌اف» برای واژگان ان‌ـگرام‌ها و واژگان «مش»

	میانگین تی‌اف‌ـ آی‌دی‌اف ان‌ـ گرام‌ها	میانگین تی‌اف‌ـ آی‌دی‌اف مش	تفاوت ان‌ـگرام‌ها با مش
گونه			
مقاله‌های چاپی	۲/۱۹	۰/۹۸	۱/۲۱
یادداشت‌ها	۲/۱۹	۱/۴۳	۰/۷۹
مقاله‌های پژوهشی	۱/۹۴	۱/۳۰	۰/۶۴
صورت‌جلسه‌های سخنرانی	۱/۷۷	۲/۲۲	-۰/۴۵
علم عامه‌پسند	۲/۰۲	۲/۶۹	-۰/۶۷
میانگین گونه	۲/۰۲	۱/۷۲	۰/۳۰
واریانس گونه	۰/۰۳	۰/۴۲	۰/۵۳
سیاق			
زبان علمی	۲/۰۹	۰/۹۲	۱/۲۷
زبان آثار چاپی	۲/۲۱	۱/۱۶	۱/۰۵
زبان علم عامه‌پسند	۱/۹۳	۱/۲	۰/۷۳
میانگین سیاق	۲/۰۷	۱/۰۹	۱/۰۲
واریانس سیاق	۰/۰۲	۰/۰۲	۰/۰۵
اصطلاح‌شناسی حوزه			
ایدز	۲/۴۳	۱/۱۳	۱/۳۷
پروتئین‌های درمانگاهی	۲/۳۲	۱/۵۰	۰/۸۲
هپاتیت	۱/۸۶	۱/۱۰	۰/۷۶
سی‌جی‌دی	۱/۷۹	۱/۸۷	-۰/۰۸
میانگین اصطلاح شناسی حوزه	۲/۱۰	۱/۴۰	۰/۷۲
واریانس اصطلاح‌شناسی حوزه	۰/۰۸	۰/۱۰	۰/۰۸
ساختار مدرک			
بحث	۲/۱۲	۱/۰۵	۱/۰۷
مقدمه	۲/۱۵	۱/۲۵	۰/۹
یافته‌ها	۲/۰۶	۱/۲	۰/۸۶
روش‌ها	۲/۰۳	۱/۷	۰/۳۳
ارجاعات	۲/۲۳	۲/۲۰	۰/۰۳
چکیده	۲/۳۶	۲/۸	-۰/۴۴
میانگین ساختار مدرک	۲/۱۶	۱/۷	۰/۴۶
واریانس ساختار مدرک	۰/۰۱	۰/۳۷	۰/۲۶
میانگین کل متغیرها	۲/۱۰	۱/۴۸	۰/۶۲
واریانس کل متغیرها	۰/۰۳	۰/۳۲	۰/۳۷

در اینجا باید به برخی از نتایج مورد انتظار اشارهٔ خاصی شود. «چکیده‌ها» و «ارجاعات» (در متغیر گفتمان «ساختار مدرک») در هر دو واژگان، بالاترین نتایج را به‌دست آوردند، اگرچه مقادیر «مش» از اهمیت بیشتری برخوردارند. این نتایج، کاربرد عام این بخش‌ها را در بازیابی اطلاعاتِ برای مدارک مرجع، تأیید می‌کنند. در رابطه با «چکیده‌ها» به عنوان توضیح می‌توان این اندیشه را ذکر کرد که در «چکیده‌ها»، اصطلاحات معنادار (واژه‌های با ارزش محتوایی بالا) از بالاترین تراکم برخوردارند. این توضیح با نظریه «لوسی ورک» (Loseeworks, 1996) هماهنگ است. «لوسی» معتقد بود که بخش‌هایی از مدرک، (مانند «چکیده») نسبت به قسمت‌های دیگر مدرک، در بردارندهٔ اصطلاحات نمایه‌ای بیشتر و بهتری هستند. به‌منظور توضیح و تبیین مقادیر «ارجاعات»، باید توجه داشت که ان‌ـگرام‌ها و «مش» معمولاً نه بر نشریات ادواری، نام نویسندگان، و مانند آن‌ها، بلکه بر «عنوان» مقاله که براستی می‌توان آن را چکیدهٔ «چکیده» تلقی کرد، تمرکز می‌کنند. بسیاری از پایگاه‌های داده‌ها، فقط اصطلاحات «عنوان» مدرک را به‌عنوان عناصر بازیابی مورد استفاده قرار می‌دهند.

[از جدول ۵] می‌توان دریافت که متغیر «ساختار مدرک» در هر دو واژگان ان‌ـگرام‌ها و «مش» تقریباً بهترین مقادیر را به دست آورده است. چنین به‌نظر می‌رسد که ساختاربندی مدارک با شیوه‌ای روشن و واضح، کمک بسیار بزرگی به امر بازیابی می‌کند.

۳. مطالعهٔ ناپایداری: با بررسی مقادیر واریانس ان‌ـگرام‌ها، می‌توان جالب‌ترین نتایج را نشان داد. نتایج به‌دست آمده نشان می‌دهند که ان‌ـگرام‌ها هنگامی‌که در یک حوزهٔ پزشکی به‌کار می‌روند و با نتایج «تی‌اف- آی‌دی‌اف» مقایسه می‌شوند، از متغیرهای گفتمان، تقریباً تأثیر نمی‌پذیرند. ارقام پایین در مقادیر درون - متغیری واریانس دیده می‌شوند: مثلاً هنگامی‌که الگوریتم [ان‌ـگرام‌ها] برای زیرمجموعه‌های «سیاق» یا هر زیرمجموعهٔ دیگری به‌کار برده می‌شود، هیچ تفاوت خاصی را در نتایج ان‌ـگرام‌ها نمی‌توان دید. مقدار کل واریانس متغیرها نیز بسیار پایین است. اما «مش»، بویژه در ارتباط با «گونه» و «ساختار مدرک»، چنین واکنشی از خود نشان نمی‌دهد. به‌نظر می‌رسد که «سیاق» تنها متغیر گفتمان است که نه بر نتایج «مش» تأثیر دارد و نه بر نتایج ان‌ـگرام‌ها. مهم‌تر آن‌که واریانس ان‌ـگرام‌ها و «مش» دقیقاً همسان هستند.

به عنوان یک نتیجهٔ جانبی، نویسندگان این مقاله چنین تصور می‌کنند که کنترل درصد اصطلاحات فیلترشده‌ای که به‌وسیلهٔ الگوریتم ان‌ـگرام‌ها فراهم شده و در واژگان «مش» یافت می‌شوند، می‌تواند میزان وابستگی «اصطلاح‌شناسی حوزه» را به دیگر متغیرهای گفتمان بسنجد: بخصوص این نکته اهمیت دارد که چگونه حوزه‌های نزدیک به هم، از مترادف‌های مختلفی برای توصیف مفهوم واحد استفاده می‌کنند. پس از این که تمام زیرمجموعه‌ها فیلتر شدند، نتایج نشان داد که از ۱۷۴۸ توصیفگر، ۴۹۸ توصیفگر که از طریق فرآیند ان‌ـگرام‌ها به‌دست آمده بودند، در واژگان «مش» یافت می‌شوند، و حدود سی درصد آن را تشکیل می‌دهند. نتایج مذکور در جدول ۶ ارائه شده است.

در رابطه با اصطلاح‌شناسی حوزه، دریافتیم که تناظر اصطلاح‌شناسی «پروتئین‌های گیاهی» در «مش»، پایین (ده درصد) است. اما وقتی مترادف‌های «مش» نیز به‌شمار آمدند، میزان تناظر حوزهٔ «پروتئین‌های گیاهی» تا ۶۳ درصد افزایش یافت. تبیین احتمالی این رفتار را می‌توان از درصد توصیفگرها در ستون (ت + م) جدول ۶ استنتاج کرد. به نظر می‌رسد که متخصصان حوزهٔ «پروتئین‌های گیاهی» همان توصیفگرهایی را که متخصصان پزشکی استفاده می‌کنند، به کار نمی‌برند. مقدار میانگین توصیفگرها که در مجموعه یافت شد، با مطالعات

پیشین بر روی حوزه‌های خاص، همخوانی داشت (Bates, 1986).

تعداد اندک اصطلاحات گردآمده از «زبان مطبوعاتی» (در متغیر «سیاق») به سنجش استنتاجات این آزمایش کمک می‌کند: «انتشارات» فقط از مجموعهٔ کوچکی از اصطلاحات پرتکرر پزشکی استفاده می‌کنند.

بعضی از نتایجی که از پیش مورد انتظار بودند، تأیید شدند. نتایج ساختار مدرک چنین بود: اصطلاحاتی که از «بخش روش‌شناسی» مدارک استخراج شدند، معمولاً در «مش» زیرسرعنوان عام «روش‌های پژوهشی» یافت می‌شدند. اصطلاحات پربسامدتر[55] در بخش‌های چکیده و نتیجه‌گیری مدارک نیز همین حالت را داشتند. تعداد زیادی از توصیفگرهایی که از بخش «ارجاع» استخراج شدند با فرانام[56]های جغرافیایی ارتباط داشتند؛ این وضع احتمالاً به این دلیل است که تمام اصطلاح‌شناسی‌های این حوزه، بیماری‌های عالم‌گیر را بازنمایی می‌کنند. این تأیید به ما امکان می‌دهد تا توجه داشته باشیم که الگوریتم‌های آینده که فقط از توصیفگرهای «مش» در زیر آن سرعنوان‌ها استفاده می‌کنند، خواهند توانست به‌صورت خودکار، مدارک «روش‌شناسی» را شناسایی کنند.

جدول ۶ . تأثیر گفتمان بر الگوریتم ان-گرام‌ها

	تعدادکل اصطلاحات به دست آمده با الگوریتم ان-گرام‌ها	درصد اصطلاحات فیلترشده منطبق با «مش»		تفاوت (ت+م)- (D)- (٪)
		مقایسه فقط با توصیفگرهای (ت) (٪) «مش»	مقایسه با توصیفگرهای «مش» و مترادفات (ت+م) (٪)	
گونه				
مقاله‌های مطبوعاتی	۵۷۹	۲۰/۶	۷۰/۳	۴۹/۷
صورتجلسه‌های کنفرانس	۳۳۰	۱۳/۹	۵۱/۸	۳۷/۹
علم عامه‌پسند	۳۴۳	۱۴/۳	۵۰/۵	۳۶/۲
مقاله‌های پژوهشی	۱۵۴۶	۱۸/۴	۵۰/۳	۳۱/۹
یادداشت‌ها	۶۸۰	۱۶/۲	۴۵/۶	۲۹/۴
سیاق				
زبان مطبوعاتی	۳۸۲	۱۵/۷	۵۲/۹	۳۷/۲
زبان علم عامه‌پسند	۸۶۷	۱۱/۲	۳۵/۸	۲۴/۶
زبان علمی	۱۵۱۵	۱۳/۹	۳۲/۰	۱۸/۱
اصطلاح‌شناسی حوزه				
پروتئین‌های گیاهی	۷۰	۱۰/۰	۶۲/۹	۵۲/۹
سی‌جی‌دی	۲۳۹	۱۳/۰	۵۲/۷	۳۹/۷
پروتئین‌های بالینی	۵۶۵	۱۳/۸	۴۳/۰	۲۹/۲
هپاتیت	۷۱۰	۱۳/۴	۴۱/۴	۲۸
ایدز	۱۶۲۲	۱۳/۰	۲۹/۹	۱۶/۹
ساختار مدرک				
بحث	۵۲۶	۳۰/۸	۸۶/۵	۵۵/۷
روش‌ها	۴۴۷	۱۴/۳	۴۴/۷	۳۰/۴
مقدمه	۸۰۴	۱۸/۸	۴۳/۷	۲۴/۹
چکیده	۳۸۷	۱۷/۱	۴۰/۱	۲۳
ارجاعات	۵۲۳	۱۵/۳	۳۲/۳	۱۷
میانگین		۱۵/۷۶	۴۸/۱۳	

۲ـ۴. آزمایش ۲: تأثیر سخن در الگوریتم کی‌ـ مینز

ساختار این آزمایش به این شرح بود:

به‌منظور دستیابی به تعداد خوشه‌هایی که باید به‌وسیلهٔ کی‌ـ مینز ساخته شوند، تعداد سلسله‌مراتب‌ها در «*مش*» محاسبه شد. ۱۵ اصطلاح (سرعنوان) در «*مش*» پیدا شد که ریشهٔ متفاوت داشتند. سطح سلسله‌مراتبی بعدی شامل ۱۱۰ اصطلاح بود.

این ۱۱۰ اصطلاح، مبنای ۱۱۰ سلسله‌مراتب «*مش*» قرار گرفتند که باید با خوشه‌های کی‌ـ مینز مقایسه و سنجیده شوند. به‌منظور مقایسهٔ خوشه‌های سلسله‌مراتبی «*مش*» با خوشه‌های الگوریتم کی‌ـ مینز، اندازه‌های زیر مشخص گردید:

- مقدار (ارزش) a: تعداد سرعنوان‌های iامین سلسله‌مراتب در «*مش*» که در jامین خوشهٔ کی‌ـ مینز حضور **نداشتند**.
- مقدار (ارزش) b: تعداد اصطلاحات jامین خوشهٔ کی‌ـ مینز که در iامین سلسله‌مراتب «*مش*» حضور **نداشتند**.

مقدار (ارزش) c: تعداد سرعنوان‌های ناامین سلسله‌مراتب «*مش*» که در لامین خوشهٔ کی‌ـ مینز حضور نداشتند.

به‌منظور مقایسهٔ نتایج، ضریب همانندی«*جاکارد*» (Romesburg, 1984) محاسبه شد:

$$c_{ij} = \frac{a}{a+b+c}$$

این ضریب باید برای هر ترکیبی از خوشه‌های کی‌ـ مینز/ *مش* محاسبه شود. چون تعداد خوشه‌هایی که قرار است ایجاد شوند باید از قبل برای کی‌ـ مینز مشخص شود، و «*مش*» نیز ۱۱۰ خوشه دارد، این الگوریتم به اجرا درمی‌آید تا چهار، شانزده، سی و دو، شصت و چهار و یکصد و بیست و هشت خوشه ایجاد کند.

به منظور مقایسه یافته‌های کلی، میانگین همهٔ ضریب‌های «*جاکارد*» به عنوان نمونه‌ای از کل یک زیرمجموعه تعیین می‌شوند. نتایج حاصله در شکل‌های بعدی ارائه گردیده‌اند. این‌ها نتایج چهار متغیر زبانی مختلف، (یعنی گونه، سیاق، اصطلاح‌شناسی حوزه، و ساختار مدرک) هستند.

تحلیل شکل‌های پیشین نشان داد که به نظر نمی‌رسد الگوریتم کی‌ـ مینز از متغیر «گونه» (شکل ۴ (الف)) و متغیر «سیاق» (شکل ۴ (ب)) تأثیر پذیرفته باشد، در حالی که متغیر «ساختار مدرک» (شکل ۴ (ج)) و متغیر «اصطلاح‌شناسی حوزه» (شکل ۴ (د)) بر آن تأثیر داشته‌اند. نمودار «۴ (ج)» نشان می‌دهد که زیرمجموعهٔ «ایدز» رفتار مناسب‌تری نسبت به دیگر زیرمجموعه‌ها داشته است. باید به‌خاطر داشت، مقادیری که «ایدز» در جدول ۵ به دست آورده، نیز بالاتر است. این شاید حاکی از وجود همبستگی میان عملکرد مناسب الگوریتم کی‌ـ مینز و «تی‌اف- آی‌دی‌اف» باشد. با توجه به این یافته‌ها، و مقایسهٔ جدول ۵ و نمودارهای مربوط به «مقاله‌های چاپی» در متغیرگونه، «چکیده‌ها» در متغیر ساختارهای مدرک و «زبان مطبوعاتی» در متغیر سیاق، تأیید می‌گردد.

از دیگر سو، هنگامی که شمار رده‌ها افزایش یافته، توصیفگرهای حاصل از بعضی اصطلاح‌شناسی‌های حوزه (نظیر «هپاتیت»)، در مقایسه با الگوریتم «*مش*» برای «ایدز» به شکل مناسبی خوشه‌بندی نشده‌اند.

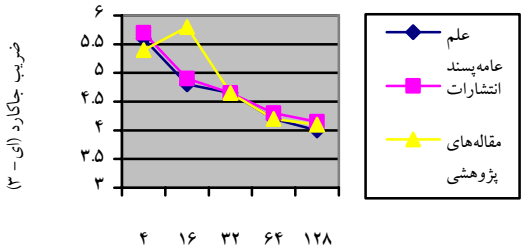
در رابطه با تأثیر اصطلاح‌شناسی حوزه در الگوریتم کی‌ـ مینز (که در شکل «۴(ج)» ارائه شده نتایج زیر را می‌توان به‌دست آورد:

این تفاوت بیش از آن که از رفتار مناسب‌تر «ایدز» ناشی شده باشد، از رفتار نامناسب‌تر «هپاتیت» و «سی‌جی‌دی» سرچشمه می‌گیرد (مقادیر حوزهٔ «ایدز» کم و بیش مانند همان نتایج «۴ (الف)» و «۴(ب)» هستند). علل احتمالی حصول چنین ارقامی می‌تواند چنین باشد: ۱) حوزهٔ «سی‌جی‌دی» حاوی

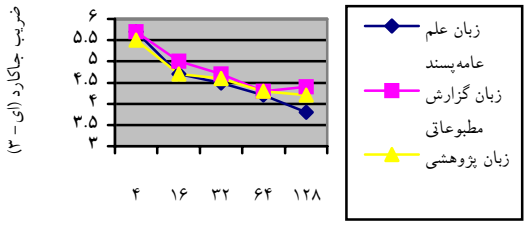
مدارک کمتری در زیرمجموعه‌ها بوده و به همین دلیل الگوریتم ان‌ـگرام‌ها بد عمل کرده، و ۲) «هپاتیت» نسبت به «ایدز»، پراکندگی خیلی بیشتری داشته و (تا سال ۱۹۹۶) در موضوع پیش‌گیری و اپیدمی‌شناسی دارای تمرکز بیشتری بوده. ۳) بیشترین تعداد توصیفگرهای «ایدز» که به وسیلهٔ الگوریتم ان‌ـگرام‌ها (نگاه کنید به جدول ۶) جمع‌آوری گردید، صراحتاً به رفتار مناسبت‌تر این الگوریتم در تعداد اندک خوشه‌ها اشاره دارد.

نمودار «۴ (د)» نشان داد که کی‌ـ مینز، مطمئناً تحت تأثیر متغیر گفتمان «ساختار مدرک» قرار داشته است. به عنوان یک مشخصهٔ کلی، مقادیر مطلق فاصلهٔ «جاکارد» در این شکل، کم‌تر از دیگر مقادیری هستند که برای دیگر متغیرهای گفتمان گردآوری شده است (مثلاً بالاترین مقدار برای «چکیده»، کمتر از ۵/۵ است). تبیین این یافته‌ها باید با این نظریه پیوند یابد که متغیر «ساختار مدرک» دلالت بر این نکته دارد که صرفاً بخش‌هایی از مدارک، به عنوان درونداد الگوریتم مورد استفاده قرار گرفته‌اند.

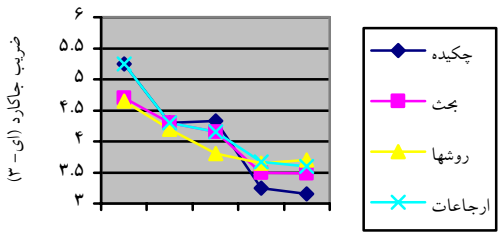
نمودار «۴(د)» همچنین حاکی از آن است که ساختار مدرک «چکیده»، هنگامی‌که توصیفگرها تعداد اندکی از خوشه‌ها را شکل داده‌اند، بالاترین نتایج را به‌دست آورده، هر چند که این پدیده، نتایج بدتری در مورد ارقام بالای خوشه‌ها داشته. این رفتار عجیب شاید با عوامل زیر ارتباط داشته باشد:



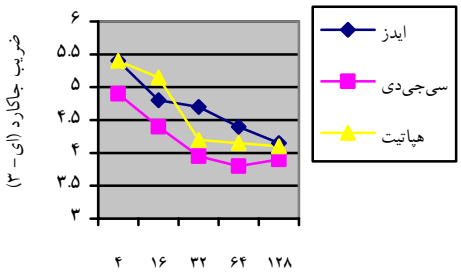
رده‌های کی- مینز الف. گونه



رده‌های کی- مینز ب.



رده‌های کی- مینز ج. ساختار سند



رده‌های کی- مینز د. اصطلاح‌شناسی حوزه

شکل ۴. الگوریتم کی‌ـ مینز در مقایسه با الگوریتم مش. محورهای افقی اشاره به تعداد رده‌های انتخاب‌شده در کی‌ـ مینز دارند. این شکل، میانگین ضریب شباهت جاکارد را با متغیرهای مختلف گفتمان نشان می‌دهد.

۱. «چکیده‌ها» دارای اطلاعات جنبی بسیار اندکی هستند، به‌نحوی که توصیفگرهای چکیده بر موضوع اصلی متن تمرکز دارند. این خصیصه بر کی‌ـ مینز تأثیر خواهد داشت. تأثیر مثبت آن هنگامی است که این الگوریتم در ساخت تعداد اندکی از دسته‌ها مؤثر واقع می‌شود و تأثیر منفی آن وقتی است که اصطلاحات می‌بایست خوشه‌های بسیاری را شکل دهند.

۲. «چکیده» معمولاً نسبت به بخش‌های دیگر مدرک، توصیفگرهای کمتری دارد (به جدول ۶ نگاه کنید). تعداد اصطلاحات به‌وضوح بر نتایج هر الگوریتم رده‌بندی اثر می‌گذارد.

می‌توان چنین نتیجه گرفت که با بهینه‌شدن تعداد خوشه‌ها، اگر این الگوریتم اصطلاحات را در تعداد رده‌های بیشتری رده‌بندی کند، نتایج حاصله معمولاً بسیار نامناسب می‌گردند.

۴ـ۳. آزمایش سوم: تأثیر گفتمان در الگوریتم چن

تحلیل با استفاده از الگوریتم چن تا حدود زیادی با روش قبلی منطبق است. هرچندکه بنا به نظر «دیاز»(Diaz et al, 1998)، هرگاه این الگوریتم برای ساخت اصطلاحنامه به‌کار رود، تفاوت میان الگوریتم کی- مینز و الگوریتم چن، به‌گونه‌ای است که چن، روابط را آشکار می‌کند؛ حال‌آن‌که الگوریتم کی- مینز، سلسله‌مراتب اصطلاحات را می‌سازد. الگوریتم چن می‌کوشد تا میزان رابطهٔ میان هر جفت واژه‌ای را که در مدارک پدیدار می‌شوند، نشان دهد.

ساختار این آزمایش این‌گونه بود: آن دسته اصطلاحاتی که با الگوریتم ان-گرام‌ها ساخته شده بودند، به‌عنوان درونداد الگوریتم چن مورد استفاده قرار گرفتند. نتیجهٔ، گروهی از روابط دو به دو میان اصطلاحات بود. این روابط با روابط موجود در «مش» مقایسه شدند.

باید این نکته را نیز مورد توجه قرار داد که سرعنوان‌های موضوعی معمول، (نظیر «مش») از سرعنوان‌های اصطلاحنامه‌ای کلاسیک، متمایزند، زیرا سرعنوان‌های معمول نسبت به سرعنوان‌های کلاسیک، روابط کمتری را در میان واژه‌ها ارائه می‌دهند. از همین‌رو روابط حاصل از الگوریتم چن تطابق اندکی را با روابط موجود در «مش» نشان می‌دهند. این پدیده می‌تواند یافته‌های جدول ۷ را تبیین کند.

جدول ۷ تعداد روابطی را که میان الگوریتم چن و «مش» مشترک هستند، بازنمایی می‌کند. این جدول نشان می‌دهد که یافته‌های بالاتر در مورد «مقاله‌های پژوهشی» متعلق به گونه، در مورد «زبان علمی» متعلق به سیاق، در مورد «سی‌جی‌دی» و «ایدز» متعلق به اصطلاح‌شناسی‌های حوزه، و در مورد «بحث» متعلق به ساختار مدرک هستند.

جدول ۷. اصطلاحات مشترک میان رده‌بندی چن و «مش»

	روابط دودویی که به‌وسیلهٔ الگوریتم چن با اصطلاحات موجود در مش به وجودآمد	درصد روابط معنایی سرعنوان یکسان مش در مقابله با کل روابط معنایی ایجادشده توسط چن
گونه		
مقاله‌های پژوهشی	۴۹۶	٪۱۱/۳
صورت‌جلسه‌های کنفرانس	۹۷	٪۹/۳
مقاله‌های علمی عامه‌پسند	۱۴۳	٪۷/۷
یادداشت‌ها	۲۱۴	٪۶/۵
اخبار	۱۱۰	٪۳/۶
واریانس		٪۶/۶
سیاق		
زبان علمی	۵۰۳	٪۱۱/۱
زبان علمی عامه‌پسند	۲۰۹	٪۷/۲
زبان مطبوعاتی	۱۴۰	٪۳/۶
واریانس		۹/۵
اصطلاح‌شناسی حوزه		
سی‌جی‌دی	۲۰	٪۱۵/۰
ایدز	۵۰۳	٪۱۱/۱
پروتئین‌های درمانگاهی	۱۲۰	٪۷/۵
هپاتیت	۲۲۲	٪۵/۴
واریانس		۱۳/۴
ساختار مدرک		
بحث	۴۵۹	٪۱۵/۰
روش‌ها	۳۸۳	٪۹/۷
مقدمه	۲۵۲	٪۹/۱
چکیده	۲۵۰	٪۸/۸
ارجاعات	۱۵۶	٪۵/۸
واریانس		٪۹/۰

از جدول ۷ می‌توان دریافت که نتایج مربوط به الگوریتم چن با آنچه از الگوریتم کی‌ـ مینز به دست آمده اشتراک اندکی دارند. در واقع، چنین به نظر می‌رسد که این دو، حداقل در متغیر گونه برای

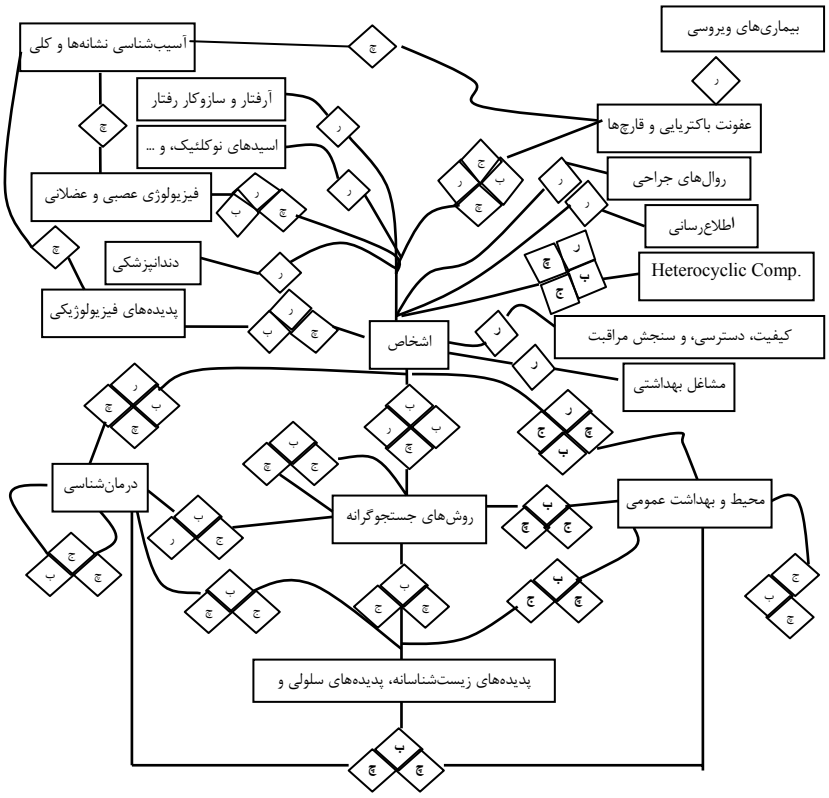
«اخبار»، در متغیر سیاق برای «زبان مطبوعاتی»، در متغیر اصطلاح‌شناسی حوزه برای «سی‌جی‌دی»، و در متغیر ساختار مدرک برای «ارجاعات» یا «چکیده»، رفتار متفاوتی داشته‌اند.

الگوریتم چن نشان می‌دهد که دو اصطلاح احتمالاً در یک مدرک واحد به هم وابستگی دارند. بنابراین هر چه مدرک کوچکتر باشد، احتمال کمتری وجود دارد که میان واژگان، وابستگی مهمی وجود داشته باشد. این نکته می‌تواند تبیین‌گر نتایج حاصل از مدارک کوچک (چون «اخبار»، «چکیده‌ها»، یا «ارجاعات») باشد.

از وابستگی اصطلاحات می‌توان برای تحلیل نحوهٔ توزیع اطلاعات در کل متن بهره برد. به هر توصیفگر با فرانامی که در «مش» دارد ارجاع داده می‌شود. بنابراین، وقوع هر جفت فرانام در بخش‌های «چکیده»، «روش‌شناسی»، «بحث» و «ارجاع»، ارزیابی می‌شود (به شکل ۵ نگاه کنید).

شکل پنج یافته‌های عینی مربوط به متغیر گفتمانی «ساختار مدرک» را نشان می‌دهد. لوزی‌هایی در میان روابط دو سرعنوان قرار دارند. حروف مندرج در لوزی‌ها هر یک ارزش خاصی دارند: حرف «چ» برای «چکیده»، حرف «ر» برای «روش‌شناسی»، حرف «ب» برای «بحث»، و حرف «ج» برای «ارجاعات» (کتابشناسی) به‌کار رفته است. هرگاه در بخش خاصی از مدرک، دو سرعنوان پیدا شده‌اند، یک لوزی مربوط به این بخش، در میان آنها قرار داده شده است.

با مطالعهٔ نتایج به‌دست آمده، گزاره‌های زیر در جدول ۸ ارائه شده‌اند.



شکل ۵. الگوریتم چن: در حوزهٔ پزشکی مورد نظر ما، میان جفت‌های اصطلاحات هر ساختار مدرک، این روابط به‌دست آمد: چ ـ چکیده، ر ـ روش‌شناسی، ب ـ بحث، ج ـ ارجاعات.

جدول ۸. توزیع وابستگی‌هایی که در الگوریتم چن برای بخش‌های مختلف مدرک یافت شد

گروه‌های بخش‌های مدرک	درصد وابستگی‌های یافت شده در تمام بخشهای گروه(٪)
«چکیده»+ «ارجاعات»+«بحث»	۲۹/۶
«روش‌شناسی»	۲۹/۶
«چکیده»+«ارجاعات»+«بحث»+«روش	۱۸/۵
«چکیده»	۱۱/۱
«چکیده»+«بحث»+«روش شناسی»	۷/۴
«ارجاعات»+«بحث»+«روش شناسی»	۳/۷

نتایج ارائه شده در جدول ۸ باید به‌شیوهٔ زیر خوانده شوند: ۲۹/۶ درصد از تمام روابطی که با الگوریتم چن به‌دست آمد، هم زمان در بخش‌های «چکیده»، «ارجاعات» و «بحث» نیز یافت شدند، درحالی‌که ۱۱/۱ درصد آن‌ها فقط در بخش «چکیده» یافت شد.

از جدول ۸ می‌توان چنین دریافت که «بحث» و «ارجاعات» غالباً (۵۲٪) با هم منطبق هستند. این شرایط احتمالاً به خاطر نمودهای بیانی گفتمان روی می‌دهند (به بخش «۲-۱» نگاه کنید). به‌موجب خصیصهٔ زبانی مناقشه‌آمیز بخش گفتمان، غالباً از ارجاعات کتابشناختی برای ادعاها استفاده می‌شود. از سوی دیگر به‌نظر نمی‌رسد که «روش‌شناسی» و «چکیده» با هم (۲۶ درصد) در لوزی‌ها ظاهر شوند.

این برآیند نشان می‌دهد که بخش «روش‌شناسی» به‌طور مناسبی در بخش چکیده بازنمایی نمی‌شود. مثلاً با یک لوزی «ر» در مورد «بیماری‌های ویروسی»، این عبارت با اصطلاح «عفونت‌های باکتریایی و قارچی» مرتبط می‌گردد، و ما از این وضعیت نتیجه می‌گیریم که هر دو اصطلاح در بخش «روش‌شناسی» مدارک یافت می‌شوند.

نتیجه‌گیری

هدف این طرح پژوهشی مطالعهٔ تأثیر متغیرهای گونه، سیاق، اصطلاح‌شناسی حوزه و ساختار مدرک بر رده‌بندی و بازیابی اطلاعات «آی‌سی‌ای‌آر»[۵۷] و الگوریتم‌های متن بود. به‌منظور دست‌یابی به این هدف باید یک چارچوب آزمایشی تعریف می‌شد تا بتوانیم با استفاده از آن، یافته‌های وابسته به گفتمان را ارزیابی کنیم و نسبت به هم بسنجیم.

معمولاً نظام‌های بازیابی اطلاعات فقط تحلیل ریخت‌شناسانه و نحوی را به‌کار می‌برند، بی‌آن‌که مشکلات مربوط به ایهام یا انسجام معنایی را حل کنند، و مطالعات اندکی دربارهٔ سنجش تأثیر گفتمان بر نظام‌های بازیابی اطلاعات انجام شده است.

کیفیت نتایج حاصله از الگوریتم‌های «آی‌سی‌ای‌آر»، بسته به متغیرهای گفتمانی مورد استفاده، متفاوت است. در آغاز انتظار می‌رفت که الگوریتم‌های تحلیل متن در علم اطلاع‌رسانی، بسته به بافتار به گونهٔ متفاوتی رفتار کنند، هرچند نتایجی که در این مقاله ارائه شد نشان می‌دهند که الگوریتم فیلترکردن ان‌ـگرام‌ها، ظاهراً تحت تأثیر متغیرهای گفتمان قرار نداشته‌اند. اما به نظر می‌رسد الگوریتم‌های رده‌بندی کی‌ـ مینز و چن تحت تأثیر این متغیرها قرار گرفته‌اند. این امر دلالت بر این نکته دارد که اگر عوامل بافتاری به‌حساب آیند، مطمئناً کارآیی این الگوریتم‌ها افزایش می‌یابد.

به‌نظر می‌آید که نوعی همبستگی را می‌توان میان «تی‌اف-آی‌دی‌اف» و الگوریتم کی‌ـ مینز نشان داد. مقادیر بالای «تی اف-آی‌دی‌اف» معمولاً دلالت بر نتایج بهتر حاصل از الگوریتم کی‌ـ مینز دارند.

این مطالعه همچنین ارزش شیوه‌ها و اصول کلاسیک علم اطلاع‌رسانی را تأیید می‌کند، شیوه‌ها و اصولی که مؤید ایده‌ای هستند که به اطلاعات

چکیده‌سازی‌شده در نمایه‌سازی مدرک، یا به ارجاعات موجود در کتاب‌سنجی اهمیت فزونتری می‌دهد (Callon et al.,1993). آرایش ساختارهای مدرک با فشردگی بالا در واژگان معنادار، برای نمایه‌سازی و در نظام‌های بازیابی اطلاعات ارزش فراوان دارد (Wormell, 1985). بعضی از ساختارها همچون چکیده‌ها، نتیجه‌گیری‌ها، شرح شکل‌ها وجدول‌ها، آغاز پاراگراف‌ها، و عنوان‌ها، در متن فراوان می‌شوند، چون طلاعات با ربط را ارائه می‌کنند (Losee, 1996)، .و بنابراین برای استخراج اطلاعات ارزشمند، می‌توان از این ساختارها بهره برد (Wormell, 1985).

در این پژوهش الگوهای کاربرد زبان را میان بیماری‌های قدیمی مانند هپاتیت، و یک بیماری‌های بسیار جدید چون ایدز مقایسه کرده‌ایم. احتمالاً دلیل فزونی تعداد اصطلاحات کنترل‌شدهٔ گوناگون در حوزهٔ هپاتیت به همین واقعیت باز می‌گردد. رفتار متفاوت الگوریتم‌های کی‌ـ مینز و ان‌ـگرام‌ها در مورد این دو نوع بیماری نیز می‌تواند با همین موضوع ارتباط داشته باشد (به شکل ۵ نگاه کنید).

منابع

Amitay, E (1998). “Using common hypertext links to identify the best phrasal description of target web documents.” **SIGIR’98 Post-conference Workshop on Hypertext IR the Web**, Melbourne, Australia.

Bates, M. (1986). “Subject access in online catalogs: a design model”. **JASIS**, 11, 357-379.

Beghtol, C. (2001). “The concept of genre and its characteristics”. **Bulletin of the American Society for Information Science and Technology**, 27(2), 17-19.

British National Corpus (http://info.ox.ac.uk/bnc/, last check 1/11/01).

Bruce, N. J. (1983). “Rhetorical constraints on information structure in medical research report writing.” **ESP in the Arab world conference**. UK: Univ. Aston.

Callon, M., Courtial, J.-P., & Penan, H. (1993). **La scientométrie**. Paris: PUF.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). “From traslation to problematic networks: an introduction to co-word analysis”. **Society Science Information**, 22, 191-235.

Chen, H., & Lynch, K. J. (1992). “Automatic construction of networks of concepts characterizing document databases”. **IEEE Transactions on systems, Man and Cybernetics,** 22, 885-902.

Cohen, J. (1995). “Highlights: Language and domain-independent automatic indexing terms for abstracting”. **JASIS,** 56(3), 162-174.

De Looze, M. A., & Lemarié, J. (1997). “Corpus relevance through Co-word analysis: an application to plant proteins”. **Scientometics**, 39(3), 267-280.

Díaz, I., Llorens, J., & Morato, J. (2002). “An algorithm for term conflation based on tree structures”. **JASIS**, 53(3), 199-208.

Díaz, I., Velasco, M., Llorens, J., & Martinez, V. (1998). “Semi-automatic construction of thesaurus applying domain analysis techniques International”. **Forum on Information and Documentation**, 23(2), 11-19.

Dijk, T. A. V. (1988). **News as Discourse. Hillsdale**, NJ: Erlbaum.

Egghe, L., & Roussau, R. (1990). **Introduction to informatics: Quantitative methods in Library, Documentation and Information Science,** Amsterdam Elsevier Science.

Frakes, W. B., & Baeza-Yates, R. (1992), **Information Retrieval: Data Structural amd Algorithms.** Upper Saddle River: Prentice Hall.

Garfield, E. (1953). “The relationship between mechanical indexing structural linguistics and information retrieval”. **Journal of Information Science**, 18, 343-354 (1992, sent to the First Sympoium on Machine Methods for Scientific Documentation (Johns Hopkins University, Mach 1953), the paper was rejected, but in 1992 the paper was published).

Gilyarevsky, R., Uzilevsky, G., & Moudrov, E. (1997). “An automatic statistical classification of different types of journal”. **International Forum on Information and Documentation**, 22(3), 24-35.

Halliday, M. A. K. (1985). **Introduction to functional grammar**. London: Arnold.

Hass, S. W., Sugarman, J., & Tibbo, H. (1996). "A text filter for the automatic identification of empirical articles". JASIS, 47(2), 167-169.

Hearst, M., & Plaunt, C. (1993). "Subtopic structuring for full-length document access". In **proceedings of the 16th ACM SIGIR conference on research and development in information retrieval**. NY: ACM.

Kando, N. (1997). "Text-level structure of research papers: Implications for text-based information processing systems". Aberdeen: British Computer Society IR SG Annual Colloquim: 1997.

Karlgren, J. (1998). "Stylistic experments for information retrieval". In T. Strzalkiwski (Ed), **Natural language information retrieval**. Tomek: Kluwer.

Karlgren, J., & Cutting, D. (1994). "Recognizing text genres with simple metrics using discriminant analysis". In: **proceeding of COLING 94**, Kyoto.

Lavid, J. (1995). "Towards a text type taxonomy: a functional ftamework for text analysis and generation". **Revista procesamiento Lenguguaje Natural**, 16, 29-43.

Lelu, C. (1997). **Modéles neuronaux pour l"analyse de données documentaires et textuelles.** Ph D Thesis, Université de Paris, Paris.

Leydesdorff, L. (1997). "Why words and cowords cannot map the development of the sciences". **JASIS**, 48(5), 418-427.

Llorens, J., Velasco, M., Martínez Orga, V. (1997). "Generación automática de representaciones de dominios. II Jornadas en Ingeniería de Software", JIS97, San Sebastián (Spain).

Llorens, J., Velasco, M., Morato, J., & Moreiro, J. A. (1998). "Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros". **Revista de Proceamiento del Lenguaje Natural**, 23, 61-68.

Losee, R. M. (1996). Text windows and phrases differing by discipline, Location in document, and sysntactic structure. **Information Processing and Management**, 32(6), 747-767.

Lowe, H. J., & Barnett G. O. (1994). "Understanding and using the medical subject headings vocabulary to perform literature searches". **JAMA**. 13, 271, No.14, pp.1103-1108.

Mitkov, R. (1998). "The latest in anaphora resolution: going multilingual". **Revista Procesamiento Lenguaje Natural**, 23, 1-7.

Morato, J. (1999). **Análisis de las relaciones cienciométricas y lingüisticas en un entorno automatizado**. Ph D thesis. Universidad Carlos III de Madrid, Leganés, Madrid (Spain).

Neighbors, J. (1981). **Software construction using components.** Ph D thesis, Department of Information and Computer Science. Irvine: University California.

Nwogu, K. N. (1997). "The medical research paper:: structure and functions". **English Specific Purposes**, 16 (2), 119-138.

Pêcheux, M. (1969). **Analyse automatique du discoure**. Paris: Dunod.

Polanco, X., Grivel, L., & Royauté, J. (1995). "How to do things with terms in informetrics: Terminological variation and stabilisation as science watch indicators". In **Proceedings of fifth international conference on scientometrics and informetrics** (pp. 435-444). Learned Information Medford (NJ).

Posterguillo, S. (1996). "Is byte popular science?" **Lenguas para finnes especificos (V). Alcalá de Henares**: Publicaciones de la Universidad de Alcalá, pp. 425-432.

Prieto-Díaz, R. (1988). "Domain analysis for reusability". In W. Tracz (Ed), **Software reuse: emetging technology** (pp. 347-353). IEEE Computer Society Press.

Romesburg, H. C. (1984). **Cluster analysis for researchers**. CA:: Lifetime Learning Publications.

Schriffrin, D. (1994). **Approaches to discourse**. Oxford: Blackwell Publishers.

Skelton, J. (1994). "Analysis of the structure of original research papers: An aid to writing original papers for publication". Journal of Documentation, 44, 455-459.

SMART project (ftp:// ftp.cs.cornell.edu/pub/smart/, last check 1/11/01).

Spark Jones, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". **Journal of Documentation**, 28, 11-21.

Swales, J. M. (1990). **Genre analysis: English in academic and research settings**. Cambridge, UK: Cambridge University Press.

Warner, A. (1994). “The role of linguistic analysis in full-text retrieval”. In **Challenges in indexing electr. Text and images** (pp. 247-264). Medford: Learned Information.

Wormell, I. (1985). **Subject Access Project (SAP)**. Improved Subject Retrieval for Monographic Publications. Ph.D. thesis, Lund: lund University.

پی‌نوشت‌ها

1. J. Moratoj, J. Lolrens, G. Genova, J. A. Moreiro (2003). "Experiments in discourse analysis impact on information classification and retrieval", *Information Processing and Management*, No. 39, pp. 825-851.
2. contextual information
3. extra-linguistic
4. n-grams algorithms
5. k-means classification
6. genre
7. register
8. Medician Subjet Headings (MeSH)
9. Co-wording
10. morphologic
11. anaphoric Situations
12. natural language processing (NLP)
13. Citing practices
14. extra-textual information
15. discourse aspects
16. document typology
17. Work-in-proress notes
18. the precision-call and pertinence ratios
19. text filter
20. alternative forms of expression
21. stylistic Variations
22. Zipf’s curve
23. field
24. tenor
25. mode of discourse
26. domain terminology
27. domain analysis (DA)
28. co-wording
29. document structure
30. co-absence
31. co-occurence
32. Academic Search Elite
33. popular-science articles
34. Blood Weekly
35. Creutzfeldt-Jakob Disease (CJD)
36. Introduction, Method, Results, Discussion (IMRD)
37. IMRD Structure
38. expanding IMRD structure
39. Term frequency-inverted document frequency (tf-idf)
40. inter-document information
41. intermediate appearance frequency
42. controlled vocabulary
43. query
44. moving center
45. co-wording algorithm
46. tem-based clusters
47. hierarchies-oriented classification algorithm
48. associations-oriented classification algorithm
49. wordnet
50. British National corpus
51. Mesh tree-structure
52. Mesh annotated alphabetic list
۵۳. نشانه هایی که بر بالای حروف گذاشته می شود مانند اعراب گذاری در زبان عربی و برخی از زبانهای دیگر.
54. weight system
55. “Heterocyclic Compounds”, “Incestigative Techniques”, and “Viruses”
56. hypernym
57. Information Classification And Retrieval (ICAR)