



## نمایه‌سازی معنایی پنهان

رسول زوارقی<sup>۱</sup>

### چکیده

این مقاله به معرفی و توصیف روش نمایه‌سازی معنایی پنهان (ال. اس. آی)<sup>۲</sup> می‌پردازد که یکی از روش‌های نوین نمایه‌سازی خودکار است. ابتدا در مقدمه‌ای کوتاه به نمایه‌سازی و چالش‌های آن اشاره می‌شود سپس در بخش دوم مدل‌های فضایی برداری که نمایه‌سازی معنایی پنهان یکی از گسترش‌های آن است، توصیف می‌شود. در بخش بعدی ضمن تشریح مفهوم نمایه‌سازی معنایی پنهان، کاربردها و موارد استفاده از آن بیان می‌گردد و سپس مبانی ریاضی آن که روش آماری تجزیه مقادیر منفرد<sup>۳</sup> است با مثالی تشریح می‌شود. در بخش بعدی فرایند کار نمایه‌سازی معنایی پنهان همراه با مثال توضیح داده می‌شود و در نهایت طرح‌ها و برنامه‌هایی که هم اکنون در این زمینه اجرا می‌شوند معرفی و به بعضی پیشرفت‌های فناورانه مؤثر در بهبود عملکرد نمایه‌سازی معنایی پنهان اشاره می‌شود.

### کلیدواژه‌ها

نمایه‌سازی معنایی، نمایه‌سازی پنهان، بازیابی اطلاعات، تجزیه مقادیر منفرد

### مقدمه

امروزه نمایه‌سازی و به‌طور کلی حوزه بازیابی اطلاعات، به سبب تغییرات قابل توجه در محیط پیرامون خود متحول شده‌اند. می‌توان بعضی از این تغییرات را این‌گونه برشمرد: گسترش قابل توجه دامنه بازیابی اطلاعات با ظهور

چندرسانه‌ای‌ها، اینترنت، و اطلاعات جهانی؛ ظهور پایگاه داده و بعضی نظریه‌های جدید در هوش مصنوعی، زبان‌شناسی و ریاضیات (۱۳؛ ۲۴؛ ۳۲)؛ رویکرد تحلیلی مباحث ریاضی و پیچیده به‌کار رفته در بازیابی اطلاعات؛ طراحی نظام‌های بازیابی اطلاعات همانند نظام‌های پایگاه

2.LSI = Latent Semantic Indexing  
3.SVD = Singular Value Decomposition

۱. کارشناس ارشد کتابداری و اطلاع‌رسانی، rasoolzavaraqi@yahoo.com

داده‌ای رابطه‌ای، و بروز امکانات جدید ناشی از فناوری‌های نوینی چون وب مانند پالایش اشتراکی<sup>۴</sup> (که با عنوان نظام‌های توصیه هدف یا شخصی شناخته می‌شود).

علاوه بر چالش‌های فوق باید به این نکته اصلی توجه داشت که در بازیابی اطلاعات مفهوم نسبتاً مبهم و گنگی به نام ربط وجود دارد که به روش‌های پیچیده تشخیص نیت کاربر و ماهیت مدارک بستگی دارد. طبق گفته پاپادیمیتریو<sup>۵</sup> و همکارانش نظریه‌های اندکی در این زمینه ارائه شده است (۳۸) که برای درک بیشتر این چالش بررسی متون کلاسیکی چون آثار ریچسبرگن و سالتون مفید خواهد بود (۴۰؛ ۴۲).

باید به یاد داشت که چالش‌های جدید، علاوه بر چالش‌های پیشین چون ذهنی‌بودن فرایند نمایه‌سازی ست که ناشی از نظری‌بودن این عمل است. اصلی‌ترین روش تعیین ذهنی‌بودن نمایه‌سازی، بررسی یکدستی آن هنگام تحلیل کار چند نمایه‌ساز از یک مدرک یا حتی یک نمایه‌ساز در زمان‌های مختلف است. همه این عوامل نشان‌دهنده پیچیدگی قابل توجه فرایند نمایه‌سازی ست که در مقاله اندرسون و پرز-کاربالو<sup>۶</sup> به طور مفصل تشریح شده است (۲؛ ۳).

وجود این چالش‌ها موجب شد که پژوهشگران از دهه ۱۹۵۰ تاکنون به دنبال روش‌های نوین و کارآمد برای رویارویی با چالش‌ها باشند و از دهه ۱۹۷۰ تاکنون برای رفع این مسئله اهتمام بسیار داشته‌اند. همان‌طور که پالگارین و گیل-لیو<sup>۷</sup> در پژوهش خود نشان دادند که بیش از ۸۰۰ اثر پژوهشی طی سال‌های ۱۹۵۶ تا ۲۰۰۰ در مورد نمایه‌سازی خودکار، نیمه‌خودکار و رایانه‌ای نگاشته شده‌اند (۳۹).

با توجه به نکات فوق، نوعی توافق عمومی وجود دارد مبنی بر اینکه بازیابی سنتی اطلاعات نمی‌تواند جوابگوی این چالش‌ها باشد. نتایج بازیابی نظام‌های سنتی بازیابی اطلاعات به دو دلیل عمده ناکارا و غیردقیق بود. اول اینکه در این نوع نظام‌ها مفهوم واحدی را می‌توان به روش‌های مختلف توصیف کرد. عبارت‌های پرسشی کاربر ممکن است در مدرک مربوط وجود نداشته باشد. دوم اینکه بیشتر کلمات بیش از یک معنا دارند در نتیجه مطابقت واژگانی عبارت‌های پرسشی کاربر ممکن است به بازیابی مدارک

نامربوط بینجامد (۱۷).

دانشمندان اصلی‌ترین علل محدودیت مدل‌ها و نظام‌های بازیابی اطلاعات را ابهام و گویانبودن<sup>۸</sup> واژه‌ها، ناکارآمدی بازنمون مدارک مجموعه، و در سویی دیگر اغتشاش<sup>۹</sup> و عدم صراحت<sup>۱۰</sup> پرسش‌های کاربر می‌دانند. راه‌حل‌های گوناگونی برای این مسائل پیشنهاد شده است که بازیابی اطلاعات براساس هستی‌شناسی یا استفاده از بازنمون معنایی مدارک و پرسش‌ها از مهم‌ترین راه‌حل‌ها هستند (۸).

در این مقاله سعی می‌شود نمایه‌سازی معنایی پنهان که نوعی بازنمون معنایی مدارک و نوع گسترش یافته مدل‌های بازیابی برداری ست و از مباحث جبر خطی و ماتریس‌های ریاضی و فن تجزیه مقادیر منفرد استفاده می‌کند، معرفی و ابعاد مختلف استفاده از آن بیان شود.

### مدل فضای برداری

از آنجا که نمایه‌سازی معنایی پنهان یکی از راهکارهایی بود که برای رفع مشکلات مدل فضای برداری به وجود آمد، در این بخش سعی می‌شود خلاصه‌ای از این مدل و مشکلات آن بیان شود.

مدل فضایی برداری یکی از چند روش تشخیص تشابه میان دو مدرک است که در سال ۱۹۷۵ به وسیله سالتون<sup>۱۱</sup> گسترش یافت (۴۳) و چارچوبی تأثیرگذار و قدرتمند برای ذخیره، تحلیل و ساختن مدارک است. این مدل در ابتدا به منظور بازیابی اطلاعات گسترش یافت و در نمایه‌سازی مدارک مبتنی بر فراوانی عبارت‌ها، به طور گسترده‌ای به کار رفت. سه مرحله این مدل عبارتند از نمایه‌سازی مدرک، وزن‌دهی عبارت، محاسبه ضریب<sup>۱۲</sup> تشابه:

۱. نمایه‌سازی مدرک: هر مدرک یا (پرسش) به صورت یک بردار در فضایی با ابعاد بزرگ‌تر ترسیم می‌شود. تعداد عبارت‌های منحصر به فرد مجموعه مدرک محاسبه می‌شود. واژگان غیرمهم از بردار مدرک حذف می‌شوند. از یک سیاهه واژگان غیرمجاز که واژگان عمومی را در خود جای می‌دهد، برای حذف واژه‌های پرکاربرد استفاده می‌شود (که در کل ۴۰ تا ۵۰ درصد از کل واژگان مدرک حذف می‌شوند).

4. Collaborative Filtering  
5. Papadimitriou  
6. Anderson Perez Carballo

7. Pulgarin & Gill-Leiva  
8. Expressionless  
9. Confusion

10. Imprecision  
11. Salton  
12. Coefficient

۲. وزن دهی عبارت: وزن دهی برای نشان دادن میزان اهمیت عبارت ها در بازنمون مدرک انجام می شود. پیش فرض اغلب روش های وزن دهی چون فراوانی معکوس مدرک<sup>۱۳</sup> این نکته است که اهمیت یک عبارت با افزایش میزان رخداد آن عبارت متناسب است. برای پیشگیری از بازیابی مدارک طولانی تر از شیوه هنجارسازی<sup>۱۴</sup> استفاده می شود (زیرا با توجه به نکته فوق، احتمال بازیابی مدارک طولانی بیشتر از احتمال بازیابی مدارک کوتاه تر خواهد بود).

۳. محاسبه ضریب تشابه: تشابه میان دو مدرک (یا میان یک پرسش و یک مدرک) با فاصله میان بردارها در فضای با ابعاد بزرگ تر تعیین می شود. بدین صورت که همپوشانی واژه، نشان دهنده تشابه خواهد بود. شناخته شده ترین مقیاس تشابه، ضریب کسینوس است و تشابه میان دو مدرک را با کسینوس زاویه میان دو بردار نشان می دهند (۱۲).

در این مدل هنگام ورود یک پرسش توسط کاربر، آن پرسش مانند دیگر مدارک فرض می شود و به صورت یک بردار نشان داده می شود. نظام در فرایند بازیابی، موقعیت پرسش را نسبت به محل هر مدرک در فضای برداری مقایسه می کند و مدارک را با توجه به میزان تشابه با پرسش کاربر رتبه بندی می کند. پس به این ترتیب یا تعداد مدارکی را که بیشترین تشابه را با پرسش دارند بازیابی می کند یا همه مدارکی را که تا حدی از آستانه تشابه (که قبلاً معین شده است) بیشتر باشد، بازیابی می کند. اصلی ترین مزیت مدل فضای برداری، رتبه بندی کارآمد و دقیق مدارک مرتبط با توجه به میزان تشابه آنها با پرسش است، در صورتی که در روش های مبتنی بر مطابقت واژگانی یا هیچ طرحی برای رتبه بندی وجود ندارد یا اگر هم وجود داشته باشد فاقد امکان رتبه بندی موارد پیچیده هستند (۵). چنانکه ممکن است به واژه ای که در اول عبارت پرسش ظاهر شده است رتبه ای بالاتر اختصاص داده شود (۱۷).

کاربران در سیاهه مدارک بازیابی شده، همیشه مدارک با بیشترین میزان تشابه را مطالعه می کنند به این دلیل که از نظر معنایی بیشتر به پرسش مربوط هستند. با وجود این سنجش میزان تشابه میان بردارها، به دستورالعملی برای چگونگی وزن دهی هر عبارت نمایه نیاز دارد. روش های مختلفی برای این امر وجود دارد که توابع دودویی<sup>۱۵</sup>، فراوانی

عبارت<sup>۱۶</sup>، و لگاریتمی<sup>۱۷</sup> بیشتر از بقیه شناخته شده اند (۱۷). البته باید توجه داشت که استفاده از این مدل مشکلاتی را نیز دارد. اصلی ترین مشکل مدل های فضای برداری، عدم مطابقت واژگان است. مدل های فضای برداری، با عبارت های نامشابه مانند اقلام نامرتبط برخورد می کنند. برای مثال رایانه و لپ تاپ<sup>۱۸</sup> اگرچه عبارت های مرتبطی هستند اما مدل های فضای برداری از کشف چنین رابطه ای عاجز هستند. پس اگر میان پرسش و مدرک در مجموعه متون هیچ کلمه مشترکی وجود نداشته باشد حتی اگر بعضی مدارک با پرسش مرتبط باشند، مقدار تشابه عملاً صفر خواهد بود و در نتیجه هیچ مدرکی بازیابی نخواهد شد. دومین اشکال این مدل ها در مجموعه مدارک بزرگ<sup>۱۹</sup> پیش می آید که در صورت تشکیل ماتریس عبارت-مدرک، ماتریس حاصل بزرگ و پراکنده خواهد بود که فضای ذخیره زیادی را می طلبد و مدت زمان پردازش و محاسبه آن نیز طولانی می شود. برای رفع این مشکلات از یکی از شاخه های این مدل با عنوان نمایه سازی معنایی پنهان استفاده می شود که در سال های اخیر رواج یافته است (۲۰؛ ۱۶؛ ۱۷).

به طور کلی می توان گفت که بازیابی اطلاعات سنتی با دو مشکل قدیمی متراف ها (مانند مدارک حذف شده مربوط به اتومبیل به هنگام پرسش درباره ماشین) و چندمعنایی ها (مثل بازیابی مدارک درباره اینترنت به هنگام پرسش درباره موج سواری) روبروست. در این مدل هیچ تفاوتی میان crane به معنای پرند ماهی خوار و crane به معنای جرثقیل نیست. برای مقابله با این دو مشکل و سایر مشکلات بازیابی سنتی اطلاعات، سعی می شود مدارک (و پرسش ها) نه با استفاده از خود عبارت ها (همان طور که در روش های برداری معمول است)، بلکه با استفاده از مفاهیم ضمنی (پنهان) آن عبارت ها، بازنمون شوند. این ساختار پنهان، نقشه ای ثابت میان عبارت ها و مفاهیم نیست بلکه به کل مدارک مجموعه و رابطه عبارت با کل آن بستگی دارد (۲۸).

### نمایه سازی معنایی پنهان

نمایه سازی معنایی پنهان یکی از فنون نمایه سازی مفهومی است که برای غلبه بر مشکلات ناشی از عدم مطابقت واژگان به وجود آمده است (۱۷). همان طور که گفته شد در

نظام‌های سنتی بازیابی، اطلاعات از مطابقت واژه به واژه عبارات‌های مدارک با پرسش کاربر بازیابی می‌شود، ولی شواهد ناکارآمدی این روش را نشان داده‌اند. همچنین از آنجا که معمولاً روش‌های متعددی برای بیان یک مفهوم وجود دارد (ترادف)<sup>۲۰</sup> احتمال دارد که میان عبارات‌ها و واژگان پرسش کاربر با بازنمون واژگانی مدرک هیچ اشتراکی وجود نداشته باشد. علاوه بر آن بیشتر واژه‌ها چندین معنا دارند (چندمعنایی)<sup>۲۱</sup> به طوری که نتیجه بازیابی با عبارات‌های پرسش کاربر، مدارک نامربوطی را دربرخواهد گرفت. رزاریو معتقد است روشی که کاربر را قادر به بازیابی اطلاعات براساس مفهوم یا معنای یک مدرک خواهد نمود و مشکلات ذکرشده را نیز برطرف می‌کند، نمایه‌سازی معنایی پنهان است (۴۱).

تعاریف مختلفی از نمایه‌سازی معنایی پنهان ارائه شده است. رزاریو آن را این‌گونه تعریف می‌کند: "نمایه‌سازی معنایی پنهان، فنی‌ست که پرسش‌ها و مدارک را به فضایی با ابعاد معنایی پنهان وارد می‌سازد"<sup>(۴۱)</sup>. پایادیمتریو و همکارانش آن را فنی برای بازیابی اطلاعات بر اساس تحلیل طیفی ماتریس عبارت مدرک تعریف می‌کنند که پیشرفت‌های تجربی پیش از آن فاقد هرگونه پیش‌بینی و توصیف محکم بود. به زعم آنها نمایه‌سازی معنایی پنهان نوعی روش بازیابی اطلاعات است که تلاش دارد تا ساختار معنایی پنهان مجموعه مدارک را با استفاده از فنون برگرفته از جبر خطی کشف کند (۳۸).

نمایه‌سازی معنایی پنهان اولین بار توسط گروهی از پژوهشگران (دیروستر و همکارانش) در بلکور<sup>۲۲</sup> در بازیابی اطلاعات به کار رفت (۱۸) و نمایه‌سازی معنایی پنهان نامیده شد. پژوهشگران به این دلیل واژه "پنهان" را به آن می‌افزایند که عبارات‌های جدید که بازنمون اطلاعات معنایی هستند، مستقیماً از مدارک یافت نمی‌شوند بلکه حاصل بررسی کل مجموعه مدارک و استفاده از روش ریاضی خاصی با عنوان تجزیه مقادیر منفرد (اس. وی. دی) است. به عقیده چنگ ساختار معنایی پنهان مدرک با توجه به الگوی استفاده از واژگان (با توجه به امکان انتخاب چندین واژه) شکل می‌گیرد. مدل نمایه‌سازی معنایی پنهان از فنون آماری خاصی برای نشان دادن ساختار پنهان معنایی و زدودن زواید

ناشی از امکان انتخاب چند واژه به جای یک مفهوم که دیروستر و همکارانش به‌طور مبسوط آن را توصیف کرده‌اند، استفاده می‌کند و با کشف الگوی استفاده از واژه‌ها، این ساختار را آشکار می‌کند و باعث حذف نوفه (پارازیت) می‌شود (۱۶). پیش فرض نمایه‌سازی معنایی پنهان این است که معمولاً کل محتوای معنایی یک متن چون پاراگراف، چکیده یا کل مدرک با مجموع معانی واژه‌های آن به‌طور تقریبی برابر است. یعنی بدین صورت که:

معنای واژه اول + معنای واژه دوم + معنای واژه سوم + ... = معنای واژه n = معنای پاراگراف  
 همچنین می‌توان با احتساب هر متن به صورت یک معادله خطی و کل مجموعه مدارک به‌صورت نظامی از معادلات هم‌زمان، معنای پایدار بازنمون‌های واژه‌ها را از کل یک مجموعه مدرک بزرگ به‌دست آورد (۳۶).

نمایه‌سازی معنایی پنهان با استفاده از بستر متن، مترادف‌ها (یعنی امکان انتخاب چند واژه به جای یک مفهوم) و چند معنایی‌ها (یعنی وجود چند معنا برای عبارتی واحد) را کنترل و بدین ترتیب از ریزش کاذب پیشگیری می‌کند. یک عبارت پنهان ممکن است مرتبط به یک مفهوم نمایان<sup>۲۳</sup> (مانند مفهوم تعامل انسان و رایانه) باشد که با چند کلیدواژه توصیف می‌شود و ترکیبی از چند واژه است (۱۲).

بنا به ادعای پژوهشگرانی چون دیروستر، دومیس، لندئر، فارناس و هارشمن<sup>۲۴</sup>، نمایه‌سازی معنایی پنهان شناخته‌شده‌ترین الگوریتم بازیابی اطلاعات است و برای مقاصد گوناگونی چون جستجو و بازیابی (۱۸)، رده‌بندی (۵۲) و پالایش<sup>۲۵</sup> (۲۱؛ ۲۳) به کار می‌رود. نمایه‌سازی معنایی پنهان یکی از روش‌های فضایی برداری برای مدل‌سازی مدارک است و طبق نظر دیروستر و دیگران؛ دومیس؛ و کنتستاتیس و پتنگر معنای "پنهان" مجموعه مدارک را آشکار می‌سازد (۱۸؛ ۲۲؛ ۳۳).

پژوهش‌های دومیس نشان دادند که نمایه‌سازی معنایی پنهان با ترجیح مفهوم معنایی مدرک بر واژگان آن، مدل‌های فضایی برداری را بهبود می‌بخشد (۲۰؛ ۲۳).

در نمایه‌سازی معنایی پنهان برخلاف مدل‌های برداری مدارک که عبارات‌ها و واژه‌ها را مستقل فرض می‌کنند، سطوح مختلفی از همبستگی<sup>۲۶</sup>، وابستگی<sup>۲۷</sup> یا پیوستگی<sup>۲۸</sup> برای

20.Synonymy                      23.Salient                              26.Correlation  
 21.Polysemy                      24.Dreewester, Dumais, Landauer, Furnas & Harshman                      27.Dependency  
 22.Bellcore                        25.Refinement                      28.Association

آنها در نظر گرفته می‌شود و این پیوستگی‌های بین عبارت‌ها با تشکیل مجموعه جدیدی از عبارت‌ها با استفاده از روش آماری تجزیه مقادیر منفرد مشخص می‌شوند (۳۷) همچنین در فضای معنایی پنهان، یک پرسش و یک مدرک، حتی در صورت نداشتن عبارت مشترک، می‌توانند تشابه کسینوسی زیادی داشته باشند زیرا عبارت‌های آنها از نظر معنایی، مشابهت مفهومی دارند در صورتی که در مدل‌های برداری چنین چیزی امکان‌پذیر نیست (۴۱).

مهمترین نکته قوت نمایه‌سازی معنایی پنهان کارآمدی بازیابی مبتنی بر پرسش کاربر است که از طریق محاسبه ماتریس حاصل می‌شود. همان‌طور که با استفاده از این روش مدارک مربوطه، حتی در صورت عدم مطابقت واژگان محتوای دو مدرک با یکدیگر، بازیابی می‌شوند (۱۲).

به‌طور کلی می‌توان گفت که نمایه‌سازی معنایی پنهان یکی از فنون روبه‌رشد نمایه‌سازی مجموعه مدارک بزرگ است که سعی دارد از یادگیری آماری ماشین<sup>۲۹</sup> در تحلیل متون استفاده کند (۳۶).

محققان مزایای مختلفی برای نمایه‌سازی معنایی پنهان برمی‌شمرند که در ادامه به چند مورد اشاره می‌شود:

چنگ، یکی از اصلی‌ترین مزایای نمایه‌سازی معنایی پنهان را استفاده از مفاهیم معنایی به جای تک‌تک واژه‌ها در نمایه‌سازی می‌داند که بدین ترتیب مدارک مربوطه حتی در صورت نبود واژه مشترک با عبارت پرسش بازیابی می‌شوند (۱۷). دپروستر نیز مزیت اصلی استفاده از بازنمون نمایه‌سازی معنایی پنهان را کارآمدی آن در رویارویی با مدارکی می‌داند که حاوی مترادف‌ها، چندمعنایی‌ها، و عبارت‌های وابسته به همدیگر است (۱۸). به نظر هازبندز، سیمون و دینگ نمایه‌سازی معنایی پنهان چون یکی از روش‌های گسترش پرسش است می‌تواند جامعیت را بهبود بخشد (۳۰).

## موارد استفاده و کاربردهای نمایه‌سازی معنایی پنهان

پژوهشگران استفاده از نمایه‌سازی معنایی پنهان را در حوزه‌های مختلف ارزیابی کرده‌اند و آن را یکی از پرکاربردترین روش‌های نمایه‌سازی دانسته‌اند.

موارد استفاده از نمایه‌سازی معنایی پنهان عبارتند از:

۱. بازیابی اطلاعات: همان‌طور که گفته شد نمایه‌سازی معنایی پنهان اولین بار توسط گروهی از پژوهشگران در بلکور در بازیابی اطلاعات به کار رفت (۱۸) و به همین نام شناخته شد. این روش بهتر از روش‌های برداری استاندارد عمل می‌کند و حتی در صورت نبود اشتراک واژگانی میان سؤال و مدرک، کارایی خود را حفظ می‌کند. استفاده از نمایه‌سازی پنهان معنایی در بازیابی اطلاعات در مقالات متعددی توسط محققانی چون بری، دومیس و ابریان؛ دومیس؛ هال<sup>۳۰</sup>؛ و هازبندز، سیمون و دینگ تأکید شده است (۹؛ ۲۳؛ ۲۹؛ ۳۰). محققان دیگری چون اندو و لی<sup>۳۱</sup>؛ بارتل، کنترل و بلو<sup>۳۲</sup>؛ دومیس؛ و ژا<sup>۳۳</sup>، مارکوس و سیمون<sup>۳۴</sup> نیز در پژوهش‌های خود نشان دادند که با استفاده از نمایه‌سازی معنایی پنهان جامعیت و مانعیت نظام بازیابی به‌طور قابل توجهی افزایش می‌یابد (۴؛ ۷؛ ۲۰؛ ۵۳). بشیری نیز از این روش در بازیابی متون فارسی استفاده کرده است (۱).

۲. بازخورد ربط: بیشتر آزمون‌هایی که از نمایه‌سازی معنایی پنهان برای بازخورد ربط استفاده می‌کنند از روشی استفاده می‌کنند که در آن، حاصل جمع برداری مدارکی که توسط کاربران مربوط شناسایی شده‌اند، جایگزین پرسش اولیه می‌شوند. پژوهش‌ها نشان داده‌اند که جایگزین‌سازی اولین مدرک مربوط با پرسش اولیه، عملکرد را در حد ۳۳ درصد و جایگزین‌سازی سه مدرک مربوط به آن عملکرد را در حد ۶۷ درصد بهبود می‌بخشد. نمایه‌سازی معنایی پنهان این قابلیت را دارد که با استفاده از فنون گسترش پرسش حتی بدون استفاده از بازخورد ربط اقدام با شناسایی مدارک مربوط کند ولی با استفاده از اطلاعات حاصل از بازخورد ربط، بازدهی عملکرد را به‌طور قابل توجهی افزایش می‌دهد.

۳. پالایش<sup>۳۵</sup> اطلاعات: استفاده از نمایه‌سازی معنایی پنهان در پالایش اطلاعات بسیار آسان است. ابتدا نمونه‌ای از مدرک با استفاده از ابزارهای استاندارد نمایه‌سازی معنایی پنهان و تجزیه مقادیر منفرد تحلیل می‌شوند و بدین ترتیب علائق کاربر به‌صورت برداری با ابعاد کم نشان داده می‌شود و سپس مدارک موجود در مجموعه با آن بردار مطابقت داده شده و در صورت تشابه مدرک با آن بردار، به کاربر توصیه

می‌شود. البته می‌توان با گذشت زمان با استفاده از روش‌های یادگیری نظیر بازخورد ربط برای بهبود بازنمون بردارهای علائق استفاده کرد (زیرا این علائق در طول زمان تغییر می‌یابند) (۴۱).

۴. همایش ارزیابی بازیابی متنی<sup>۳۶</sup>: اخیراً از نمایه‌سازی معنایی پنهان در زمینه‌های پالایش و بازیابی اطلاعات همایش ارزیابی بازیابی متنی استفاده می‌شود. پرسش‌های این برنامه بسیار طولانی و حاوی توصیفات مفصلی هستند که گاه طول آنها به ۵۰ کلمه نیز می‌رسد. پرسش‌های این برنامه، دارای ابزارهای قوی‌تری نسبت به محاسن نمایه‌سازی معنایی پنهان یا سایر روش‌هایی که سعی در غنی‌سازی پرسش‌های کاربران دارند، است. اصلی‌ترین چالش این مجموعه، گسترش ابزارهای نمایه‌سازی معنایی پنهان برای کنترل حجم مجموعه بود که این نتایج، کاملاً دلگرم‌کننده بودند. از آنجا که در زمان همایش‌های ارزیابی بازیابی متنی محاسبه کل مدارک مجموعه منطقی نبود، از نمونه‌ای در حدود ۷۰ هزار مدرک و ۹۰ هزار عبارت استفاده شد. چنین ماتریس‌های عبارت-مدرکی بسیار پراکنده هستند و فقط ۰/۰۰۱ تا ۰/۰۰۲ درصد ورودی‌ها را دربر می‌گیرند، برای مثال محاسبه دویستمین مقدار فردی بزرگ‌تر، به حدود ۱۸ ساعت زمان کار واحد پردازش مرکزی<sup>۳۷</sup> در یک پایگاه سانس پارک<sup>۳۸</sup> با ده ایستگاه کاری نیاز داشت. با وجود دشواری بسیار مقایسه تفصیلی نظام‌ها (به علت تفاوت‌های قابل توجه پیش‌پردازش، بازنمون و مطابقت)، عملکرد نمایه‌سازی معنایی پنهان بسیار خوب گزارش شد (۲۱). استفاده از اطلاعات درباره مدارک شناخته شده مربوط برای ایجاد برای هر پرسش، در پالایش بسیار سودمند بود. مزیت بازیافت ۳۱ درصدی تا حدی کمتر از میزان مشاهده شده در سایر آزمون‌های پالایش بود که به پرسش‌های اولیه این همایش نسبت داده می‌شود. در بازیابی نیز با استفاده از نمایه‌سازی معنایی پنهان در مقایسه با روش‌های برداری کلیدواژه‌ای ۱۶ درصد بهبود مشاهده شد (۹).

۵. بازیابی در متون چندزبانه: چون نمایه‌سازی معنایی پنهان از دستور زبان انگلیسی استفاده نمی‌کند می‌توان از آن در هر زبانی استفاده کرد. به علاوه می‌توان از آن در بازیابی از متونی که چندزبانه هستند استفاده نمود و پرسش‌های

کاربران (در هر زبان موجود در متون) با مطابقت زبان پرسش با مدارک هم‌زبان صورت می‌گیرد. لندوئر و لیتمن<sup>۳۹</sup> روشی را برای ایجاد فضایی مشترک که واژه‌های هر زبان در آن ارائه شوند با استفاده از نمایه‌سازی معنایی پنهان توصیف می‌کنند (۳۵). در این نوع موارد ماتریس اولیه عبارت-مدرک، با استفاده از مجموعه‌ای از چکیده‌های چندین (که در پژوهش آنها فرانسوی و انگلیسی بود) تشکیل می‌شود و به هر چکیده به صورت ترکیبی از نسخه‌های انگلیسی و فرانسوی، نگاه می‌شود. تجزیه مقادیر منفرد کاهش یافته، ماتریس عبارت-چکیده ترکیبی محاسبه می‌شود. فضای حاصل از چکیده‌های ترکیبی انگلیسی و فرانسوی تشکیل می‌شود. بدین ترتیب در این فضای کاهش یافته، واژه‌های انگلیسی و فرانسوی که در چکیده‌های ترکیبی مشابه ظاهر می‌شوند، در کنار همدیگر قرار خواهند گرفت. بعد از این تحلیل، چکیده‌های تک‌زبانی وارد می‌شوند؛ یک چکیده فرانسوی به سادگی در بردار حاصل جمع واژه‌های اجزای اصلی که قبلاً در فضای نمایه‌سازی معنایی پنهان وارد شده بودند قرار می‌گیرد. پرسش‌های انگلیسی یا فرانسوی با چکیده‌های انگلیسی یا فرانسوی مطابقت داده می‌شوند. در فضای نمایه‌سازی معنایی پنهان نیازی به ترجمه نیست. تجربه‌ها نشان داده‌اند که فضای چندزبانی کاملاً خودکار، حتی بهتر از فضایی با زبان واحد عمل می‌کند و نتایج بهتری عاید کاربران می‌کند. عملکرد بازیابی مدارک فرانسوی زبان با پرسش‌های به زبان انگلیسی (و برعکس) برابر با عملکردی است که ابتدا پرسش‌ها به فرانسوی ترجمه و سپس در پایگاه داده منحصرأ فرانسوی جستجو شوند. این روش نتایج خوبی را در بازیابی چکیده‌های انگلیسی و اندیشه نگاشت‌های زاپنی کنجی و ترجمه‌های چندزبانی (انگلیسی و یونانی) ارائه داد (۵۱).

۶. مطابقت افراد به جای مدارک: از دیگر کاربردهای نمایه‌سازی معنایی پنهان یافتن افراد خبره در یک حوزه خاص با استفاده از مقالات و آثار آنهاست. در این کاربرد، اشخاص براساس مقالاتی که نوشته‌اند معرفی می‌شوند. برای مثال در پژوهش فرناس<sup>۴۰</sup> و دیگران که بلکور ادوایز<sup>۴۱</sup> نام دارد نظامی برای یافتن خبرگان محلی باتوجه به پرسش‌های کاربران طراحی شد. یک پرسش با جدیدترین مدارک مطابقت

36.TREC = Textual Retrieval Evaluation Conference  
37.CPU = Central Processing Unit  
38.SUNSPARC

39. Landauer & Littman  
40.Furnas  
41.Bellcore Advisor



داده می‌شد و توصیف و مشخصات مربوط به پدیدآوران به عنوان مربوط‌ترین مطلب ارائه می‌شد (۲۷). در موردی دیگر، از نمایه‌سازی معنایی پنهان برای تخصیص مقالات به افراد به منظور ارزیابی داوری آنها استفاده شد. برای انجام این کار ابتدا صدها منتقد بر اساس متون تألیفیشان توصیف شدند که این کار بنیان تحلیل نمایه‌سازی معنایی پنهان شد. صدها مقاله نیز بر اساس چکیده آنها نشان داده شدند و با نزدیک‌ترین و مرتبط‌ترین داوران مطابقت داده شدند. استفاده از امکان تشابه‌یابی این برنامه برای تخصیص مقالات همایش<sup>۴۲</sup> تعامل انسان- رایانه<sup>۴۳</sup> به داوران انجام شد. تحلیل‌های بعدی نشان داد که این کار کاملاً خودکار، به خوبی کاری بود که توسط افراد خبره انجام شده بود (۹).

۷. پرسش‌ها و مدارک اشتباه: نمایه‌سازی معنایی پنهان به دلیل عدم وابستگی به مطابقت کلیدواژه‌ای، در رویارویی با دروندادهای اشتباه، که در نمودخوان نوری<sup>۴۲</sup>، درونداد باز<sup>۴۳</sup>، یا خطاهای هجی کردن رخ می‌دهد، بسیار مفید است. در صورت بروز خطا در هنگام پوشش<sup>۴۴</sup> و هجی غلط واژه‌های چون Dumais که Duniais نوشته شود، سایر واژه‌های مدرک، صحیح هجی خواهند شد و در صورتی که واژه‌هایی که به صورت صحیح هجی شده‌اند در مدارکی ظاهر شوند که نسخه صحیح هجی شده Dumais را داشته باشند، قرار Dumais در فضای کاهش یافته، نزدیک Duniais خواهد گرفت (۹).

۸. سایر موارد: اسچانتز<sup>۴۵</sup>؛ و گالانت<sup>۴۶</sup> از تجزیه مقادیر منفرد و ایده‌های مرتبط به کاهش ابعاد در ابهام‌زدایی معنایی واژه و کار بازیابی اطلاعات استفاده کردند (۲۸؛ ۴۵). هال؛ و یانگ و چوت<sup>۴۷</sup> از تجزیه مقادیر منفرد و نمایه‌سازی معنایی پنهان و در اولین مرحله ارتباط با رده‌بندی آماری استفاده کردند. چنانکه با استفاده از ابعاد مأخوذ از این نمایه‌سازی، تعداد متغیرهای پیشگوی<sup>۴۸</sup> رده‌بندی به‌طور قابل توجهی کاهش یافتند (۲۹؛ ۵۰). وو<sup>۴۹</sup> و همکارانش نیز از این برنامه در کاهش ابعاد مجموعه آموزش نظام رده‌بندی پروتئین شبکه عصبی که در ژنوم<sup>۵۰</sup> انسان به کار می‌رود استفاده کردند (۴۸). پژوهش پایادیمتریو و همکارانش نشان داد که فور؛

فلیکتر و دیگران؛ جلیف؛ و کرن و دیگران نیز در پژوهش‌های خود از این روش استفاده کرده‌اند (۳۸؛ ۲۶؛ ۲۵؛ ۳۱؛ ۳۴). رزاریو<sup>۵۱</sup> یکی از کاربردهای عمده این نمایه‌سازی را سنجش میزان تشابه می‌داند و آن را جایگزینی برای مقیاس‌های سنتی مبتنی بر همپوشانی واژه‌ها چون TF.IDF می‌داند (۴۱). پژوهش‌های بیکر و مکلام<sup>۵۲</sup>؛ دومیس؛ و یانگ نشان داد که این برنامه نمایه‌سازی دسته‌بندی متن را بهبود می‌بخشد (۶؛ ۲۳؛ ۴۹) و پژوهش اسچانتز نشان داد که استفاده از آن در ابهام‌زدایی مفهوم واژه خیلی سودمند است (۴۴). هازبندز، سیمون و دینگ در پژوهش خود نشان دادند علاوه بر اینکه استفاده از نمایه‌سازی معنایی پنهان در مجموعه مدارک کوچک سودمند و اثرگذار است، اثربخشی آن در مجموعه مدارک بزرگ بیشتر است. آنها در پایان با ذکر اینکه استفاده از این روش مانعیت بازیابی را افزایش نمی‌دهد با پیشنهاد روش خاصی برای عادی‌سازی عبارت، این نقصان را برطرف کردند (۳۰). لازم به ذکر است که محققانی چون بارتل<sup>۵۳</sup> و همکارانش؛ دینگ<sup>۵۴</sup>؛ پایادیمتریو و همکارانش؛ و ژا و همکارانش؛ قبلاً سودمندی استفاده از این برنامه نمایه‌سازی را در مجموعه مدارک کوچک نشان داده بودند (۷؛ ۱۹؛ ۳۸). و در نهایت پایادیمتریو و همکارانش نیز ثابت کردند که این برنامه در شرایط خاص در اقتباس معنای ضمنی مجموعه مدارک بزرگ موفق عمل می‌کند و عملکرد بازیابی را بهبود می‌بخشد. آنها سازوکاری با عنوان پیش‌بینی تصادفی را برای تسریع نمایه‌سازی معنایی پنهان پیشنهاد کردند (۳۸).

سایر موارد استفاده از نمایه‌سازی معنایی پنهان عبارتند از تحلیل تشابه عمومی (۱۴؛ ۱۵)، طرح استارواکر<sup>۵۵</sup> (۱۶). پرتز و همکارانش نیز از آن در پیش‌بینی روند گسترش فناوری‌ها استفاده کردند (۵۵).

### تجزیه مقادیر منفرد

نمایه‌سازی معنایی پنهان بر نوعی فن ریاضی با عنوان "تجزیه مقادیر منفرد" استوار است که مبانی جبری آن برای اولین بار در مقاله دیروستر و همکارانش توصیف شد (۱۸) و

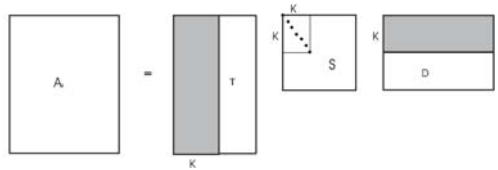
42. OCR = Optical Character Reader  
43. Open  
44. Scan  
45. Schutz  
46. Gallant

47. Yang & Chute  
48. Predictor Variables  
49. Wu  
50. Genome  
51. Rosario

52. Baker & Maccallum  
53. Bartell  
54. Ding  
55. Starwalker

بعدها در مقالات بری، دومیس و ابراین<sup>۵۶</sup> بیشتر به آن پرداخته شد (۹). این مقالات فرایند این فن را توصیف و ماتریس‌های حاصل را از نظر هندسی تفسیر کردند. مقاله وایمر-هستینگز<sup>۵۷</sup> نیز نشان داد که توانایی و کارایی قابل توجه نمایه‌سازی معنایی پنهان ناشی از روش ریاضی "تجزیه مقادیر منفرد" است (۳۳؛ ۴۷). لازم به ذکر است که از این سازوکار در مجموعه‌های بیش از نیم میلیون مدرک حاوی ۷۵۰ هزار نوع واژه منحصر به فرد برای سنجش تشابه دو مدرک استفاده می‌شود (۳۶).

کنستانتیس و پتنگر<sup>۵۸</sup> درباره رابطه میان عملکرد نمایه‌سازی معنایی پنهان و مقادیر ماتریس عبارت-مدرک بحث کردند و نتیجه گرفتند که این برنامه بر تفاوت‌های معنایی (معانی پنهان) تأکید می‌کند در نتیجه نوفه (پارازیت) را در داده‌ها کاهش می‌دهد (۳۳). در این نمایه‌سازی برخلاف مدل فضایی برداری سنتی که مدارک و پرسش‌ها به صورت بردارهایی در فضای  $t$  بعدی (تعداد عبارت‌های نمایه شده مجموعه است) ارائه می‌شدند از تجزیه مقادیر منفرد برای تجزیه ماتریس عبارت-مدرک به سه ماتریس استفاده می‌شود. این سه ماتریس عبارتند از  $T$  که یک ماتریس عبارت



-اندازه است،  $S$  یک ماتریس مقدار فردی (اندازه-اندازه) و  $D$  یک ماتریس مدرک-اندازه. تعداد اندازه‌ها نیز که رتبه ماتریس عبارت-مدرک است با  $r$  نشان داده می‌شود. در یک نظام نمایه‌سازی معنایی پنهان، ماتریس‌های  $T$ ،  $S$  و  $D$  به میزان  $k$  کاهش می‌یابند (۳۳).

در تصویر فوق (که برگرفته از منبع ۹ است و نحوه کاهش ماتریس را با استفاده از تجزیه مقادیر منفرد توضیح می‌دهد) نواحی هاشور خورده ماتریس‌های  $T$ ،  $D$ ، باقی می‌مانند چون بیشترین مقدار فردی را دارند و نواحی غیرهاشور خورده نیز حذف می‌شوند. هدف از کاهش اندازه در فرایند تجزیه مقادیر منفرد و نمایه‌سازی معنایی پنهان، کاهش نوفه (پارازیت)

در فضای پنهان است که باعث می‌شود تا ساختار رابطه واژگانی غنی‌تر، معانی پنهان کنونی مجموعه را آشکار سازد (۳۳).

نمایه‌سازی معنایی پنهان برای کاهش اندازه از پارامتر  $k$  استفاده می‌کند که این مقدار برای هر مجموعه به صورت تجربی، محاسبه می‌شود. به‌طور کلی به دلیل هزینه محاسبه الگوریتم تجزیه مقادیر منفرد و ذخیره و مقایسه بردارهایی با ابعاد بزرگ، مقادیر کوچک برای  $k$  ترجیح داده می‌شوند (۳۳).

پژوهشگرانی چون ژا و سیمون نیز نمایه‌سازی معنایی پنهان را با استفاده از مدل‌های فضایی فرعی توصیف کردند و آزمونی آماری برای انتخاب میزان بهینه ابعاد یک مجموعه پیشنهاد نمودند (۵۴). استوری<sup>۵۹</sup> درباره رابطه آن با روش‌های رگرسیون آماری و قضیه بیس بحث کرده است (۴۶). دینگ نیز یک مدل آماری برای سنجش تشابه کسینوس ارائه داد (۱۹). اسپاتز نیز قابلیت‌های قابل توجه تجزیه مقادیر منفرد را مطالعه کرده است (۴۵).

از آنجا که ساخت ماتریس تجزیه مقادیر منفرد از نظر محاسباتی گران است تمهیداتی اندیشیده شده که اندازه ماتریس به صورت قابل توجهی کاهش یابد. این کاهش اندازه علاوه بر اینکه هزینه محاسباتی را کاهش می‌دهد بلکه به هنگام بازیابی، نوفه (پارازیت) را به‌طور قابل توجهی کاهش می‌دهد و مدارک را به‌طور خودکار، به صورت ساختارهای معنایی متناسب با بازیابی اطلاعات سازماندهی می‌کند (۱۲). انتخاب ابعاد، مسئله مهمی در نمایه‌سازی معنایی پنهان است به‌گونه‌ای که کاهش متناسب ابعاد، نوفه‌ها و بازیابی‌های ناخواسته را کاهش می‌دهد و در مقابل کاهش زیاد تعداد ابعاد باعث از میان رفتن اطلاعات مهم در هنگام بازیابی می‌شود (۴۱).

حال در این بخش برای روشن شدن هر چه بیشتر مطلب، فن ریاضی تجزیه مقادیر منفرد با مقدمه و مثال توصیف می‌شود. ماتریس  $F$  را که به صورت  $m \times n$  است در نظر بگیرید که از دو بعد تشکیل شده است. اگر  $m=n$  باشد  $F$  را مربع می‌گویند. در ماتریس معکوس<sup>۶۰</sup> این ماتریس که با  $F^{-1}$  نشان داده می‌شود چنین ویژگی وجود دارد که  $FF^{-1}=F^{-1}F$ . در صورتی که یک ماتریس مربع، ماتریس

56. Berry, Dumais & O'Brien  
57. Wiemer-Hastings  
58. Kontostathis & Pottenger  
59. Story  
60. Inverse



معکوس داشته باشد به آن غیرمنفرد<sup>۶۱</sup> و در غیر این صورت به آن منفرد<sup>۶۲</sup> گویند. مقلوب<sup>۶۳</sup> ماتریس F نیز که با F<sup>t</sup> نشان داده می‌شود با تعویض جای سطرها و ستون‌ها با هم رخ می‌دهد و یک ماتریس در صورتی ماتریس واحد<sup>۶۴</sup> نامیده می‌شود که مقلوب و معکوس آن یکسان باشد یعنی FF<sup>-1</sup>=FF<sup>t</sup> با نام F در صورتی رتبه r را به خود اختصاص می‌دهد که بزرگ‌ترین زیرماتریس مربع غیرمنفرد F یک ماتریس r×n باشد. با استفاده از تجزیه مقادیر منفرد می‌توان ماتریس قراردادی F را با ابعاد m×n این چنین نشان داد  $F_v = A^{1/2} U^t$  که U, V ماتریس‌های

$$A^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & & & \vdots & & \\ & \ddots & & \vdots & & 0 \\ & & \lambda_r^{1/2} & \vdots & & \\ \dots & \dots & \dots & \vdots & \dots & \\ & & 0 & \vdots & & 0 \end{bmatrix}$$

واحد m×n، n×n، n×n هستند و A<sup>1/2</sup> یک ماتریس m×n با قطر<sup>t</sup> 1/2 است که یک مقدار منفرد F نامیده می‌شود. از آنجا که U, V ماتریس‌های واحد هستند می‌توانیم بنویسیم که  $F = UA^{1/2} V^t$  ستون‌های ماتریس واحد U از بردارهای ویژه<sup>۶۵</sup> ماتریس متناسب<sup>t</sup> FF<sup>t</sup> و ستون‌های ماتریس واحد V، از بردارهای ویژه ماتریس متناسب<sup>t</sup> F<sup>t</sup>F تشکیل یافته‌اند. بنابراین می‌توان

$$F = \sum_{j=1}^r \lambda_j^{1/2} u_j v_j^t$$

ماتریس F را برحسب بردارهای ویژه این چنین نشان داد: حاصل ضرب بردارهای ویژه بیرونی فوق، مجموعه‌ای از ماتریس‌های با رتبه واحد را تشکیل می‌دهند که هر یک از

$$F = \begin{bmatrix} 6 & 6 \\ 0 & 1 \\ 4 & 0 \\ 0 & 6 \end{bmatrix}$$

آنها با مقدار منفرد مربوط F آنها مشخص می‌شوند (۳۷). مثال برای تجزیه مقادیر منفرد: برای محاسبه تجزیه مقادیر منفرد ماتریس F بدین ترتیب عمل می‌کنیم: مرحله اول: ابتدا V را محاسبه می‌کنیم:

$$F'F = \begin{bmatrix} 52 & 36 \\ 36 & 73 \end{bmatrix}$$

مقادیر ویژه ماتریس فوق به ترتیب ۲۵ و ۱۰۰ هستند و بردارهای ویژه آنها نیز به ترتیب و [۰, ۸, -۶, ۰]<sup>t</sup>، [۰, ۶۰, ۸, ۰]<sup>t</sup> هستند.

مرحله دوم: U را محاسبه می‌کنیم:

$$FF' = \begin{bmatrix} 72 & 6 & 24 & 36 \\ 6 & 1 & 0 & 6 \\ 24 & 0 & 16 & 0 \\ 36 & 6 & 0 & 36 \end{bmatrix}$$

این ماتریس فقط دو مقدار ویژه غیرصفر دارد و بردارهای ویژه آن

$$[0, 840, 080, 240, 48]^t$$

$$[0, 24, 0, 120, 64, 0, 72]^t$$

هستند.

مرحله سوم F: را در شکل تجزیه مقادیر منفرد بدین نحو نشان می‌دهیم:

$$F = (100)^{1/2} [0, 48, 080, 240, 48]^t$$

$$[0, 60, 8]^t + (25)^{1/2}$$

$$[0, 24, 0, 120, 64, 0, 72]^t [0, 8, 0, 6]^t$$

### شیوه کار نمایه‌سازی معنایی پنهان

حال پس از روشن شدن روش کار تجزیه مقادیر منفرد که درک آن لازمه درک مراحل کار نمایه‌سازی معنایی پنهان است در این بخش به صورت مفصل و با مثالی مراحل کار این نمایه‌سازی توصیف می‌شود.

برنر، چن و بویاک روش کار را چنین بیان می‌کنند (۱۲):

61. Nonsingular  
62. Singular  
63. Transpose

64. Unitary  
65. Eigenvector

system engineering testing of EPS  
 D5: Relation of user-perceived response time to error measurements  
 D6: Generatinrag ndom, binary, unordered trees  
 D7:The intersection graph of paths in trees  
 D8: Graph minors: widths of trees and well-quasi-ordering  
 D9: Graph minors: a survey  
 واژه‌هایی که در بیش از یک مدرک ظاهر شده‌اند با قلم سیاه نشان داده شده‌اند و در نمایه‌سازی و ساخت ماتریس پیوستگی عبارت-مدرک به کار می‌روند. این ماتریس نشان می‌دهد که هر عبارت چند بار در مدرک تکرار شده است.

عبارت	D1	D2	D3	D4	D5	D6	D7	D8	D9
human	۱	*	*	۱	*	*	*	*	*
interface	۱	*	۱	*	*	*	*	*	*
computer	۱	۱	*	*	*	*	*	*	*
user	*	۱	۱	*	۱	*	*	*	*
system	*	۱	۱	۲	*	*	*	*	*
response	*	۱	*	*	۱	*	*	*	*
time	*	۱	*	*	۱	*	*	*	*
EPS	*	*	۱	۱	*	*	*	*	*
survey	*	۱	*	*	*	*	*	*	۱
trees	*	*	*	*	*	۱	۱	۱	*
graph	*	*	*	*	*	*	۱	۱	۱
minors	*	*	*	*	*	*	*	۱	۱

حال در این مرحله و با تشکیل ماتریس عبارت مدرک درون‌داد مدل تجزیه مقادیر منفرد فراهم شد. یکی از جالب‌ترین جنبه‌های رویکرد این مدل این است که می‌توان با آن عبارت و مدرک را در یک فضا به صورت نقاطی مانند حاصل ضرب داخلی (مقیاس کسینوس) میان نقاط مشاهده کرد. با تخمین ویژگی<sup>۶۷</sup> تجزیه مقادیر منفرد، اندازه فضای بازنمون کاهش می‌یابد. در عمل فضا طوری انتخاب شده است که برای مقادیر ۱۰۰ تا ۲۰۰ مدرک که حاوی چند هزار عبارت باشند پاسخگو باشد.

در نمودار زیر بازنمون دوبعدی همه مدارک و عبارت‌های مثال فوق نشان داده شده است. در این بازنمون دو سطر اول ماتریس  $UA^{1/2}$ ، مختصات عبارت‌ها و همچنین دو سطر

۱. نمونه‌های بازنمون مجموعه مدارک به ماتریس واژه‌های عنوان/پدیدآور/چکیده مقالات وارد می‌شوند. ورودی‌های سلول‌های ماتریس، فراوانی واژه در عنوان/پدیدآور/چکیده یک مدرک هستند.

۲. در این مرحله اطلاعات حاصل از وزن‌دهی مدارک به تجزیه مقادیر منفرد وارد می‌شوند.

۳. با استفاده از تجزیه مقادیر منفرد یک فضای معنایی چکیده  $n$  بعدی ساخته شده و هر واژه به صورت یک بردار نشان داده می‌شود.

۴. برآیند بردارهای واژه‌های محتوایی آن مدرک بدون احتساب ترتیب آنها، بازنمون نهایی نمایه‌سازی معنایی پنهان از مدرک محسوب می‌شود (۱۲).

به طور ساده می‌توان اصلی‌ترین مراحل کاری نظام بازیابی اطلاعات در نمایه‌سازی معنایی پنهان را موارد زیر دانست:

۱. کاهش خودکار عبارت‌های مجموعه مدارک،  
 ۲. به دست آوردن مدارک از پایگاه‌های داده ذخیره شده در آن،

۳. بیش فرآوری مجموعه مدارک با استفاده از عبارت‌ها و حذف واژگان غیرمجاز<sup>۶۶</sup>،

۴. ساخت ماتریس عبارت-مدرک،  
 ۵. محاسبه میزان تجزیه مقادیر منفرد ماتریس عبارت-مدرک،

۶. بازیابی مدارک بر اساس پرسش کاربر.

در این بخش رویکرد نمایه‌سازی معنایی پنهان با استفاده از یک مثال توصیف می‌شود.

مجموعه کوچکی از مدارک فرضی D1 تا D9 را در نظر بگیرید. این مدارک در دو موضوع متفاوت قرار دارند. ۵ مدرک اولی در زمینه تعامل انسان- رایانه و چهار مدرک مابقی در زمینه گراف‌ها هستند.

D1: Human machine interface

for computer applications in finance

D2: A survey of user

opinion of computer system response time

D3: The EPS user interface

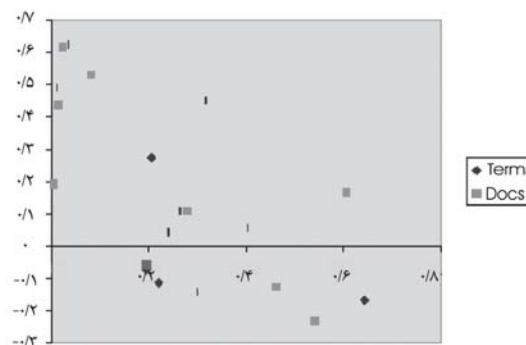
management system

D4: System and human

66. Stop words

67. Property

اول  $VA^{1/2}$  نیز به عنوان مختصات مدارک در نظر گرفته



برای استفاده از این بازنمون در بازیابی، عبارت‌های پرسش در فضای فوق ترسیم و مدارک کاذبی ساخته می‌شود. برای مثال اگر پرسش Human-Computer Interaction

باشد مدارک کاذب در نقطه‌ای بین عبارت‌های Human و Computer قرار می‌گیرد و در مرحله بعد با توجه به سنجش متناسب با مدارک موجود در فضای مدارک مقایسه می‌شود (در صورت نیاز). این رویکرد امکان کشف پیوستگی‌های پیچیده و غیرمستقیم میان عبارت‌ها و مدارک را به کاربر می‌دهد (۳۷).

### پیشرفت‌های فناورانه و چشم‌انداز نمایه‌سازی معنایی پنهان

رزاریو معتقد است که امروزه بعضی پیشرفت‌های رایانه‌ای هستند که استفاده از نمایه‌سازی معنایی پنهان را به‌ویژه در مجموعه‌های بزرگ‌تر سودمندتر می‌سازند (۴۰). مثلاً:

۱. محاسبه کارآمد تجزیه مقادیر منفرد کاهش یافته در ماتریس‌های بی‌نهایت بزرگ و پراکنده؛
۲. روزآمدسازی تجزیه مقادیر منفرد بی‌درنگ در پایگاه‌های داده‌ای که به‌طور مکرر تغییر می‌کنند؛

۳. مطابقت کارآمد پرسش‌ها با مدارک (یعنی یافتن مدارک و پرسش‌های نزدیک به هم در فضایی با ابعاد زیاد) (۴۱).

مبحث نمایه‌سازی معنایی پنهان با توجه به نو بودن و داشتن قابلیت‌های فراوان مورد توجه پژوهشگران مختلف قرار گرفته است چنانکه مراکز پژوهشی خاصی به دنبال بهبود و گسترش آن هستند. در ایران نیز بشیری در رساله کارشناسی ارشد خود به این روش پرداخت و با استفاده از مدل مجموعه‌های نامنظم میزان عملکرد آن را به‌صورت عملی و تجربی در بازیابی متون فارسی ارزیابی کرد و به نتایج قابل توجهی دست یافت به‌نحوی که مانند سایر پژوهش‌های مشابه در خارج از کشور به قابلیت‌های خاص این روش در بازیابی اطلاعات اذعان داشت (۱). در ادامه چند منبع وب شناخته شده نمایه‌سازی معنایی پنهان در سطح جهان معرفی می‌کنند:

۱. صفحه نمایه‌سازی معنایی پنهان تلکردیا<sup>۶۸</sup> (بلکور سابق) به نشانی:

<http://lsi.research.telcordia.com>

۲. وب سایتی در دانشگاه کلرادو به نشانی

**نمایه‌سازی معنایی پنهان یکی از گام‌های عملی جامعه اطلاع‌رسانی برای تحقق نمایه‌سازی خودکار است که علاوه بر کاربرد در بازیابی اطلاعات، در زمینه‌های مختلفی چون بازخورد ربط، پالایش اطلاعات، بازیابی متنی، بازیابی در متون چندزبانه، مطابقت افراد به جای مدارک، بازیابی متونی که اشکالات چاپی دارند، ابهام‌زدایی معانی واژه، رده‌بندی آماری، و سنجش میزان تشابه استفاده می‌شود**

متون فارسی کار شده است که نتایج قابل توجهی در برداشت که امید می رود در زمینه سایر کارکردهای این روش نمایه سازی در زبان فارسی نیز پژوهش شود.

### منابع

۱. بشیری، حسن. «بررسی مدل فازی و ارزیابی نمایه سازی معنایی پنهان در بازیابی متون فارسی». پایان نامه کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر دانشگاه شهید بهشتی، ۱۳۸۳.

2. Anderson, J. D.; Perez-Carballo, J. "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing". *Information Processing and Management*, No. 37 (2001): 231–254.

3. Ibid. "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort". *Information Processing and Management*, No. 37 (2001): 255–277.

4. Ando, R.; Lee, L. "Iterative residual rescaling: An analysis and generalization of LSI". In *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval*, 2001, pp: 154–162

5. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. [s.l.]: Addison-Wesley, 1999.

6. Baker, L.; McCallum, A. "Distributional

<http://www.lsi.research.telcordia.com>

۳. وب سایتی در دانشگاه تنسی به نشانی

<http://www.cs.utk.edu/~lsi>

۴. SVDPackc (نسخه ۱/۰) که توسط مایکل بری<sup>۶۹</sup>

توسعه یافته است و شامل چهار روش عددی (تکراری) برای محاسبه مقادیر فردی ماتریس های پراکنده بزرگ با استفاده از مانعیت مضاعف ANSI Fortran<sup>۷۰</sup> است، به نشانی <http://netlib.org/svdpack>

۵. تجزیه گر عمومی متن<sup>۷۰</sup> که توسط هوارد، تنگ، بری و مارتین<sup>۷۱</sup> در دانشگاه تنسی طراحی شده است که یک بسته نرم افزاری یکپارچه شی گرا (C++) برای ایجاد ساختارهای داده ها و کدگذاری لازم برای مدل های بازیابی اطلاعات است. به نشانی:

[/http://www.cs.utk.edu/~lsoft.html](http://www.cs.utk.edu/~lsoft.html)

### حاصل سخن

همان طور که در این مقاله عنوان شد روش نمایه سازی معنایی پنهان یکی از گام های عملی جامعه اطلاع رسانی برای تحقق نمایه سازی خودکار است که علاوه بر کاربرد در بازیابی اطلاعات، در زمینه های مختلفی چون بازخورد ربط، پالایش اطلاعات، بازیابی متنی، بازیابی در متون چندزبانه، مطابقت افراد به جای مدارک، بازیابی متونی که اشکالات چاپی دارند، ابهام زدایی معانی واژه، رده بندی آماری، و سنجش میزان تشابه استفاده می شود. این روش در کارکردهای خود قابلیت های خاصی دارد که مانع از بروز ریزش کاذب، نوفه و اختلال در بازیابی اطلاعات می شود و اصولاً برای انجام کارکرد خود نیازی به ساختار دستوری و نحوی یک زبان ندارد و حتی قادر به بازیابی اطلاعات غیرمتنی چون جدول ها، تصاویر نیز هست که این قابلیت ها در سایر روش های نمایه سازی نادر هستند. از دیگر نقاط قوت این روش باز بودن جای بحث و پژوهش در این زمینه است که مراکز پژوهشی مختلفی در دانشگاه های معتبر جهان چون تلکریدا، کلرادو، و تنسی به پژوهش در زمینه رویکردها و ابعاد مختلف این روش نمایه سازی مشغول هستند. همان طور که مطرح شد در ایران نیز در زمینه استفاده از این روش در بازیابی

69. Michael Berry

70. GTP = General Text Parser

71. Howard, Tang, Berry & Martin

13. Brewer, E. *Invited talk*, PODS-SIGMOD, 1997.
14. Chen, C. "Tracking latent domain structures: An integration of Pathfinder and Latent Semantic Analysis". *AI & Society*, Vol. 11, No. 1-2 (1997): 48-62
15. Ibid. "Visualizing semantic spaces and author co-citation networks in digital libraries". *Information Processing and Management*, Vol. 35, No. 2 (1999): 401-420.
16. Chen, C. M.; Paul, R. J. "Visualizing a knowledge domain's intellectual structure". *IEEE Computer*, Vol. 34, No. 3 (2001): 65-71. [on-line]. Available: <http://www.pages.drexel.edu/~cc345/papers/computer.html>.
17. Cheng, B. "Towards Understanding Latent Semantic Indexing". 2005. [on-line]. Available: <http://www.cs.ualberta.ca/TechReports/2003/TR03-03/TR03-03.pdf>.
18. Deerwester, S. C. ... [et al]. "Indexing by latent semantic analysis". *Journal of the American Society of Information Science*, Vol. 41, No. 6 (1990): 391-407. [on-line]. Available: <http://www.citeseer.nj.nec.com/deerwester90indexing.html>.
19. Ding, C. H. Q. "A similarity-based probability model for latent semantic indexing". In *Proceedings of the twenty-second annual international ACM/SIGIR conference on research and development in information retrieval*, 1999, pp: 59-65.
20. Ibid. "Improving the retrieval of information clustering of words for text classification". In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval*, 1998.
7. Bartell, G. C.; Cottrell, G. W.; Belew, R. "Representing documents using an explicit model of their similarities". *Journal of the American Society for Information Science*, No. 46 (1995): 251-271.
8. Baziz, M.; Boughanem, M.; Aussenac-Gilles, N. "Semantic networks for a conceptual indexing of documents in IR". *Presented in 7th ISPS Algeters May 2005*.
9. Berry, M. W.; Dumais, S. T.; O'Brien, G. W. "Using linear algebra for intelligent information retrieval". *SIAM Review*, Vol. 37, No. 4 (1995): 575-595. [on-line]. Available: <http://www.citeseer.nj.nec.com/berry95using.html>.
10. Borner, K. "Extracting and visualizing semantic structures in retrieval results for browsing". *Paper presented at the ACM Digital Libraries, San Antonio, Texas, (27 June 2000): 234-235*.
11. Borner, K.; Dillon, A.; Dolinsky, M. "LVis - Digital Library Visualizer". *Paper presented at the Information Visualisation 2000, Symposium on Digital Libraries, London, England, 2000, pp: 77-81*.
12. Borner, K; Chen, C.; Boyack, K. "Visualizing knowledge domains". *Annual review of information and technology*, Vol. 37 (2003).

word sense disambiguation using neural networks”. *Neural Computation*, No. 3 (1991).

29.Hull, D. “Improving text retrieval for the routing problem using latent semantic indexing”. In *Proceedings of the 17th ACM Conference on Research and Development in Information Retrieval*, 1994, pp: 282-290.

30.Husbands, P.; Simon, H.; Ding, C. “Term norm distribution and its effects on latent semantic indexing”. *Information Processing and Management*, No. 4 (2005): 777-787.

31.Joliffe, I. T. *Principal Component Analysis*. New York: Springer-Verlag, 1986.

32.Kleinberg, J. “Authoritative sources in a hyperlinked environment”, In *Proceedings ACM-SIAM Symposium on Discrete Algorithms*, 1998.

33.Kontosthathis, A.; Pottenger, W. M. “A framework for understanding Latent Semantic Indexing (LSI) performance”. *Information Processing and Management*, No. 42 (2006): 56–73.

34.Korn, F. ... [et al]. “Ratio rules: a new paradigm for fast, quantifiable data mining”, In *Proceedings. VLDB*, 1998, pp: 582-593.

35.Landauer, T. K.; Littman, M. L. “Fully automatic cross-language document retrieval using latent semantic indexing”, In *Proceedings of the 6th Annual Conference*

Research Methods”, *Instruments, Computers*, Vol. 23, No. 2 (1991): 229-236.

21.Ibid. “Latent semantic indexing (LSI) and TREC-2”. In D. Harman (Ed.), *The Second Text Retrieval Conference (TREC-2)* (1994): 105–116. [on-line]. Available: [citeseer.nj.nec.com/19248.html](http://citeseer.nj.nec.com/19248.html).

22.Dumais, S. T. “LSI meets TREC: A status report”. In D. Harman (Ed.), *The First Text Retrieval Conference (TREC-1)* (1992):137–152. [on-line]. Available: [citeseer.nj.nec.com/dumais93lsi.html](http://citeseer.nj.nec.com/dumais93lsi.html).

23.Ibid. “Using LSI for information filtering: TREC-3 experiments”. In D. Harman (Ed.), *The Third Text Retrieval Conference (TREC-3)* (1995): 219–230.

24.Fagin, R. “Combining fuzzy information from multiple sources”, In *Proceedings. 1996 PODS*, 1996, pp: 216-223.

25.Flickner, M. ... [et al]. “Query by image and video content: The QBIC system”. *IEEE Computer*, No. 28 (1995): 23-32.

26.Fuhr, N. “Probabilistic models of information retrieval”. *IEEE Computer*. No. 35 (1992): 244-255.

27.Furnas, G.W. ... [et al]. “Information retrieval using a singular value decomposition model of latent semantic structure”, In *Proceedings of SIGIR*, 1988.

28.Gallant, I. A. “practical approach for representing contexts and for performing



tor space model for automatic indexing”. *Communications of the ACM*, Vol. 18, No. 11 (1975): 613-620.

45.Ibid. “Automatic word sense discrimination”. *Computational Linguistics*, Vol.

44.Schutze, H. “Dimensions of meaning”. *In Proceedings of supercomputing*, 1992. 24, No. 1 (1998): 97-124.

46.Story, R. E. “An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model”. *Information Processing and Management*, Vol. 32, No. 3 (1996): 329-344.

47.Wiemer-Hastings, P. “How latent is latent semantic analysis?” *In Proceedings of the sixteenth international joint conference on artificial intelligence*, 1999, pp: 932-937.

48.Wu ... [et al]. “Neural Networks for full-scale protein sequence classification: Sequence encoding with SVD”. Machine Learning, 1994. [Research project].

49.Yang, Y. “An evaluation of statistical approaches to text categorization”. *Journal for Information Retrieval*, No. 1 (1999): 9-90.

50.Yang Y.; Chute, C. G. “An application of least square fit mapping to text information retrieval”. *In Proceedings of ACM-SIGIR Conference*, 1993.

51.Young, P.G. “Cross-Language Information Retrieval Using Latent Semantic Indexing”. Master’s thesis, The University of

of the UW Centre for the New Oxford English Dictionary and Text Research, UW Centre for the New OED and Text Research, Waterloo Ontario, 1990.

36.Landuer, T. K.; Laham, D.; Derr, M. “Collaquium paper mapping knowledge domains: from paragraph to graph: latent semantic analysis for information visualization”. *Proceedings National of Academic Science USA*, Vol. 101, Supplement 1 (6 Apr. 2004): 5214–5219.

37.“Latent Semantic Indexing”. [on-line]. Available: <http://lsi. argreenhouse.com/lsi/LSIpapers.html>

38.Papadimitriou, C. ... [et al]. “Latent semantic indexing: A probabilistic analysis”. *Journal of Computer and System Sciences*, No. 61 (2000): 217-235.

39.Pulgarin, A.; Gil-Leiva, I. “Bibliometric analysis of the automatic indexing literature: 1956–2000”. *Information Processing and Management*, No. 40 (2004): 365–377.

40.Rijsbergen, C. J. Van. *Information retrieval*. London: Butterworth, 1979.

41.Rosario, B. “Latent semantic indexing: an overview”. *INFOSYS 240*, 2000. [on-line]. Available: <http://www.cse.msu.edu/~cse960/Papers/LSI/LSI.pdf>.

42.Salton, G.; McGill, M. *Introduction to Modern Information Retrieval*. New York: Mac-Graw Hill, 1983.

43.Salton, G.; Yang, C.; Wong, A. “A vec-

thirteenth symposium on the interface, 1998, pp: 315–320.

54.Zha, H.; Marques, O.; Simon, H. “A subspace-based model for information retrieval with applications in latent semantic indexing”. In *Proceedings of the Irregular 98*, Lecture notes in computer science, Vol. 1457 (1998): 29-42.

55.Zhu, D.; Porter, A. L. *Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change*, 2002.

Knoxville, TN, 1994.

52.Zelikovitz, S.; Hirsh, H. “Using LSI for text classification in the presence of background text”. In H. Paques, L. Liu, & D. Grossman (Eds.), *Proceedings of CIKM-01, tenth ACM international conference on information and knowledge management*, 2001, pp: 113–118. New York: ACM Press. [on-line]. Available: <http://www.citeseer.nj.nec.com/zelikovitz01using.html>.

53.Zha, H.; Simon, H.. “A subspace-based model for latent semantic indexing in information retrieval”. In *Proceedings of the*