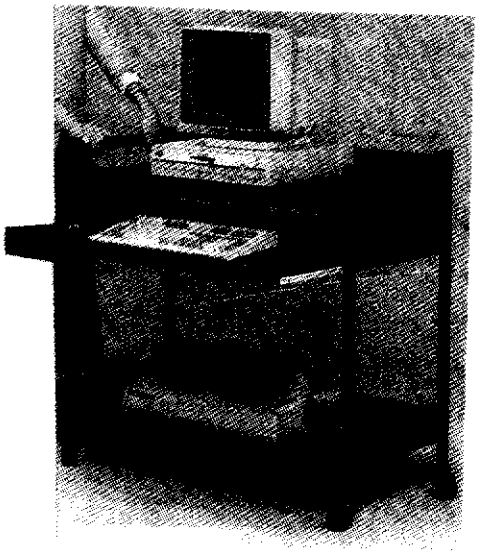
نمایه‌سازی خودکار

گذشته، حال، آینده

نویسنده: عباس گیلوری
عضو هیأت علمی مرکز اطلاع‌رسانی و خدمات علمی وزارت جهاد کشاورزی



مقدمه

پیش از این، اطلاعات موجود بر روی رسانه‌هایی چون کتاب، پایگاه‌های اطلاعاتی و انواع ریزنمون‌ها مانند ریزفیلم[1] و ریزبرگه[2] ذخیره شده‌اند. امروزه، با گسترش و توسعه اطلاعات الکترونیکی از یک سو و رشد سریع و آنی اطلاعات اینترنتی از سوی دیگر، بر حجم اطلاعات قابل‌دسترس افزوده شده است. برای دسترس‌پذیری اطلاعات (چه به‌صورت ماشین‌خوان و چه به‌صورت دستی)، متخصصین اطلاع‌رسانی ناگزیرند از رایانه و برنامه‌های رایانه‌ای بهره گیرند. حجم اطلاعات موجود، تنوع ساختار اطلاعات در قالب‌هایی چون اصل مقاله، خلاصه مقالات و چکیده و ضرورت یکسان‌سازی قالب آنها از یک طرف و ارائه درخواست‌های اطلاعاتی بهره‌گیران به‌صورت اصطلاح یا واژه‌های منفرد ازطرف دیگر باعث شده است تا ایجاد نمایه‌ها **و انطباق درخواست‌ها با شیوه‌هایی دستی بسیار دشوار گردد.**

در طول سی سال گذشته و با تلاش‌های متخصصین اطلاع‌رسانی به‌نظر می‌رسد که نمایه‌سازی خودکار می‌تواند به‌عنوان ابزاری برای ساماندهی و در دسترس‌قراردادن حجم عظیم اطلاعات موجود یا به‌عبارت بهتر مقابله با انفجار اطلاعات مطرح گردد. با افزایش هر روزه انتشارات الکترونیکی، نمایه‌سازی و کاوش‌های تمام‌متن در بسیاری از برنامه‌های بازیابی اطلاعات توسعه فراوانی داشته است(۱). به‌نظر می‌رسد که در بسیاری از برنامه‌های بازیابی اطلاعات می‌توان فرآیندهای الگوریتمی رایانه‌ای را برای پردازش‌های اطلاعاتی و به‌عنوان جایگزینی برای پردازش‌های فکری بشری بکار گرفت. در فرآیندهایی چون تولید نمایه‌های خودکار، طبقه‌بندی مدارک و اصطلاحات، تدوین راهبردهای کاوش و یا ایجاد ارتباط بین اصطلاحات وابسته می‌توان از رایانه بهره گرفت. اگرچه تاکنون از رایانه در اغلب فعالیت‌های بازیابی همچون استخراج پیام از متن مدارک، تدوین اصطلاحنامه‌ها، فهم نیازهای اطلاعاتی بهره‌گیران از طریق درخواست‌های آنها و ... استفاده شده است اما تاکنون رایانه نتوانسته است در هیچ موضوعی به اندازه حوزه نمایه‌سازی خودکار به‌صورت مستقل انجام وظیفه نماید(۲).

نمایه

در واژه‌نامه "اصطلاحات نمایه‌سازی"، نمایه این‌گونه تعریف شده است :

"فهرستی از اصطلاحات که به روشی قابل‌فهم و واضح مرتب شده و محلی که اطلاعات در آن وجود دارند را نشان می‌دهد"(۳).

در کتاب "نمایه سازی و چکیده‌نویسی"، آقای راولی تعریف دیگری از نمایه را ارائه کرده است. او می‌گوید: "نمایه، مجموعه سازمان‌یافته‌ای از نقاط دسترسی[3] است که بهره‌گیری از اطلاعات معلوم به مجموعه‌ای از اطلاعات ناشناخته و نامعلوم را رهنمون می‌سازد"(۴). از دو تعریف یادشده فوق، اولین تعریف ناظر بر مجموعه‌های کاغذی است و زمانی ارائه شده که هنوز نمایه‌های ماشینی و پایگاه‌های اطلاعاتی چندان کاربرد عام نداشته‌اند. دومین تعریف با عنایت به ورود رایانه در حوزه کتابداری ارائه شده و به‌نظر می‌رسد از جامعیت

بیشتری برخوردار باشد. بدیهی است که نمایه نقش بسیار مهمی در بازیابی اطلاعات داشته و نمایه‌ساز که به تحلیل محتوای اطلاعاتی مدارک می‌پردازد، در ارتباط و انتقال اطلاعات نقش بارزی دارد. ایجاد نمایه از طریق انجام نمایه‌سازی میسر است. نمایه‌سازی را "ثبت و ضبط محتوای اطلاعاتی مدارک با استفاده از کلیه روش‌ها و دستورالعمل‌ها به‌منظور سازمان‌دادن اطلاعات و به قصد سهولت بازیابی"(۵) تعریف کرده‌اند.

امروزه نمایه‌ها از اصلی‌ترین اجزاء پایگاه‌های اطلاعاتی به‌شمار می‌آیند و به‌طور گسترده در ایجاد پایگاه اطلاعاتی - چه بر روی رایانه‌ها و چه به‌صورت درون‌خطی و بر روی اینترنت - مورداستفاده قرار می‌گیرند. نمایه‌ها عموماً سه کارکرد عمده را بر عهده دارند:

- محتوای اطلاعاتی مدارک را فشرده می‌سازند.
- به‌عنوان واسطه‌ای برای تطبیق و یکسان‌سازی زبان مدرک و زبان کاوش بکار می‌روند.
- به‌عنوان ابزاری کارا بر شیوه تدوین راهبردهای کاوش در جستجوهای اطلاعاتی نظارت دارند(۶).

زبان‌های نمایه‌سازی

زبان نمایه‌سازی، زبانی است که برای شرح موضوع یا سایر جنبه‌های اطلاعات یا مدارک در یک نمایه بکار می‌رود. درواقع، زبان‌های نمایه‌سازی بر شیوه گزینش یا استخراج کلمات نمایه (چه از درون متن و چه از طریق واژگان کنترل‌شده)دلالت دارند. زبان‌ها نمایه‌سازی بسیار متفاوتند اما آنها را می‌توان به سه دسته اصلی تقسیم کرد.

زبان‌های نمایه‌سازی کنترل شده[۴]

در این روش، واژه‌ها یا اصطلاحاتی به مدرک مورد نمایه منتسب می‌شوند. واژه‌های انتسابی از مجموعه‌ای از واژگان کنترل شده گرفته می‌شوند. درواقع، در نمایه‌سازی کنترل‌شده، واژگان نمایه عیناً از درون متن گرفته نمی‌شود بلکه واژه نویسنده متن با فهرست اصطلاحات مقابله شده و اصطلاح پذیرفته شده انتخاب می‌شود.

زبان‌های نمایه‌سازی آزاد[۵]

برخلاف زبان‌های نمایه‌سازی کنترل‌شده، این زبان‌ها به‌راحتی محدود نمی‌شوند. در این زبان‌ها، هر واژه‌ای که برای موضوع مدرک مناسب باشد به‌عنوان اصطلاح نمایه‌ای انتخاب می‌شود. زبان نمایه‌سازی آزاد در محیط‌های نمایه‌سازی رایانه‌ای بسیار متداول است.

نمایه‌سازی به زبان طبیعی[۶]

در این روش، برای نمایه‌سازی از زبان خود مدرک کمک گرفته می‌شود. به‌طور کلی می‌توان گفت که نمایه‌سازی به زبان طبیعی نوعی از زبان نمایه‌سازی آزاد است. نمایه‌سازی به زبان طبیعی بیشتر با تعیین رایانه‌ای اصطلاحات سروکار دارد و به زبان عنوان، چکیده و سایر اجزاء متن به‌عنوان یک رکورد اطلاعاتی وابسته است(۴). در این روش انتخاب اصطلاحات ساده است و به گزینش و تحلیل مدارک نیازی نیست. نکته قابل توجه آنکه هنوز درباره اینکه آیا نمایه‌سازی به زبان طبیعی می‌تواند ما را به بازیابی کارا رهنمون سازد یا خیر بحث وجود دارد.

نمایه‌سازی رایانه‌ای یا ماشینی

"فرآیند استخراج مجموعه‌ای از مدخل‌های نمایه‌ای که بیانگر موضوع متن هستند توسط رایانه از متن ماشین‌خوان را نمایه‌سازی خودکار می‌نامند"(۷). در پایگاه‌های اطلاعاتی از رایانه به شیوه‌های مختلفی برای نمایه‌سازی استفاده می‌شود. نمایه‌سازی پایگاه‌های اطلاعاتی و نمایه‌های ماشینی به روش‌های زیر ممکن است انجام گیرد:

۱. استفاده از رایانه برای انجام امور دفتری نمایه‌سازی مثل ورود اطلاعات در پایگاه‌های اطلاعاتی؛

۲. بهره‌گیری از رایانه برای کنترل کیفیت نمایه‌های تولیدی. مثلاً بررسی این مسئله که آیا همه اصطلاحات نمایه در اصطلاحنامه وجود دارند یا خیر؟؛

۳.بهره‌گیری فکری از رایانه مثل استفاده از رایانه. برای مثال استفاده از رایانه برای وزن‌دهی و انتخاب اصطلاحات نمایه؛

۴. نمایه‌سازی کاملاً خودکار به کمک رایانه (۸).

امروزه تولیدکنندگان پایگاه‌های اطلاعاتی در سطوح مختلف برای نمایه‌سازی از رایانه بهره می‌گیرند. درحال حاضر، نمایه‌سازی خودکار رایانه‌ای برای نمایه‌سازی بخشی از پایگاه اطلاعاتی مثل موضوعات خاص، نقاط دسترسی یا نوع مدرک مورداستفاده قرار می‌گیرد. مرکز اطلاع‌رسانی فن‌آوری دفاعی[7] برای مدیریت متون در پایگاه اطلاعاتی خود، مرکز FI2 Karlsruhe در نمایه‌سازی نام‌های شیمیایی، مرکز بین‌المللی اطلاعات علمی و فنی روسیه[8] برای منابع زبانی و Inspec برای نمایه‌سازی فایل‌های پشتیبانِ استانداردهای جدید خود از نمایه‌سازی خودکار بهره گرفته‌اند.

نکته‌ای که در نمایه‌سازی ماشینی باید یادآور شد آن است که ماشین‌ها و ابزارهای نمایه‌سازی فعلی صرفاً به‌منظور نمایه‌سازی مدارک ابداع نشده‌اند بلکه این ماشین‌ها اصولاً برای انجام امور تجاری مانند کارهای آماری و حسابداری‌های کلان طراحی شده و برای نمایه‌سازی با اصول و معیارهای آن منطبق شده‌اند(۹). به‌عبارت دیگر،محدودیت‌های این ماشین‌ها برای اهداف نمایه‌سازی عمدتاً بر این حقیقت استوارند که آنها به‌جای طراحی برای نمایه‌سازی، با مباحث نمایه‌سازی منطبق شده‌اند. بنابراین، امکانات آتی نمایه‌سازی ماشینی شاید بیشتر از آنی باشد که مخترعین این ابزارها از آن در ذهن دارند.

نمایه‌سازی به کمک رایانه[9]

نکته‌ای که در اینجا باید به آن اشاره داشت اینکه در امر نمایه‌سازی خودکار باید بین دو اصطلاح نمایه‌سازی به‌کمک رایانه و نمایه‌سازی خودکار[10] تفاوت قائل شد. گرچه در بسیاری از متون تخصصی بین این دو اصطلاح تفاوت قائل نشده‌اند اما باید گفت که در نمایه‌سازی خودکار پس از تدوین برنامه رایانه‌ای و تدارک الگوریتم‌های خاص، نیروی انسانی در فرآیند تولید نمایه هیچ دخالتی ندارد. اما در نمایه‌سازی به کمک رایانه، نمایه‌ساز از رایانه برای انجام امور نمایه‌سازی استفاده می‌کند(۱۰). اولین نمایه‌های رایانه‌ای مانند کوئیک و کووک از نوع نمایه‌سازی به کمک رایانه بوده‌اند.

نمایه‌سازی خودکار

با توسعه روزافزون اینترنت و حجم گسترده اطلاعات درون‌خطی و ناپیوسته و همچنین افزایش متن‌های الکترونیکی و نشر سنتی، دسترسی به مطالب و منابع علمی موردنیاز را دشوار ساخته است. به‌نظر می‌رسد که بدون نمایه‌سازی خودکار به‌ویژه بر روی اینترنت، دسترسی به منابع بسیار دشوار باشد. "نمایه‌سازی خودکار فرآیندی است که در آن مجموعه‌ای از اصطلاحات که بیانگر موضوع مدرک هستند توسط رایانه استخراج شده یا همه کلمات متن مدرک در یک فایل مغلوب قرار می‌گیرند"(۱۱). ابداع نمایه‌سازی خودکار که عموماًبر بهره‌گیری و استفاده از اصطلاحات واحد تأکید دارد، با اختراع رایانه‌های بزرگ هم‌زمان بود. اولین نمایه‌سازی خودکار که توسط اچ پی لون[11] انجام گرفت بر بسامد واژه‌ها در نمایه‌سازی متون تأکید داشت. این تحقیقات ادامه یافت و اکنون به روش‌های نمایه‌سازی خودکار پیچیده‌ای انجامیده است که بر استفاده از اصطلاحات واحد[12] و استفاده از بخش بیشتری از اطلاعات مانند عبارت تأکید دارند(۱۲).

نکته‌ای که در نمایه‌سازی خودکار باید موردتوجه قرار گیرد آن است که هر نوع نمایه‌سازی خودکار باید دو فرآیند پایه را موردنظر قرار دهد: اول، باید به شیوه‌ای واحدهای متنی را مجزا نماید. این عمل مستلزم فهم زبان طبیعی است. دوم، بر اساس واحدهای متنی یک زبان، کلمات نمایه را گزینش و استخراج کند. باید متذکر شد که به‌دلیل تنگناهای زبانی در زبان‌های نمایه‌سازی موجود، بهره‌گیری از ابزارهای نمایه‌سازی مثل اصطلاحنامه یا نظام رده‌بندی چندان مفید نیست(۱۱).

عوامل مؤثر در نمایه‌سازی خودکار

در ایجاد نمایه خودکار باید به عواملی توجه نماییم، توجه به این عوامل ما را در تهیه نمایه‌ای کارا کمک خواهد نمود. این عوامل عبارتنداز:

۱. محدوده رکورد. اولین تصمیم‌گیری مهم برای تهیه هر نوع نمایه، گزینش حد و حدود رکوردی است که واحد قابل‌جستجو را تعریف می‌کند. طول رکورد می‌تواند یک کتاب، فصلی از کتاب، بخشی از فصل یا حتی یک پارگراف

باشد. این تصمیم‌گیری در بازیابی کارآمد نقشی حیاتی دارد.

اگر مدرک در حال نمایه‌سازی مثلاً مقاله‌ای از یک مجله است، بدیهی است که طول رکورد، تمام متن مقاله خواهد بود. اما این شیوه تعریف فیلد معمولاً در هنگام استفاده در پایگاه‌های اطلاعاتی، کاوش‌های اطلاع‌رسانی را با دشواری مواجه می‌سازد. بدیهی است که نباید کل مقاله یا کتاب را به عنوان یک رکورد درنظر گرفت. هر چه طول مدارک در حال نمایه‌سازی بزرگتر باشد از دقت نمایه‌سازی کاسته شده و امر نمایه‌سازی خودکار نیز دشوارتر می‌گردد. بنابراین، نباید کل مدرک را به‌عنوان یک رکورد اطلاعاتی تعریف کرد. همچنین نباید طول رکورد نیز بیش از اندازه کوتاه باشد چرا که حجم کم متن رکورد، بازیابی ناقص را در پی دارد و ضعف بازیابی به نوبه خود از ضعف نمایه ناشی می‌شود(۱۲). ازاین رو، تعیین یک جمله از متن به‌عنوان رکورد چندان مناسب نیست. بنابراین، در نظر گرفتن یک پارگراف به‌عنوان رکورد مناسب به‌نظر می‌رسد.

۲. انتخاب اصطلاحات نمایه

تعیین حد و حدود یک واژه از دیگر مسائلی است که در نمایه‌سازی خودکار باید به آن توجه شود. در نظام‌های نمایه‌سازی دستی، گزینش کلمات برای نمایه به سهولت انجام می‌شود. اما در نمایه‌سازی خودکار از آنجا که ماشین از هوشمندی لازم برای انتخاب کلمات برخوردار نیست بنابراین باید حدود کلمه را تعریف کرد. معمولاً حدود کلمات نمایه را با استفاده از علائم نقطه‌گذاری تعریف می‌کنند. به‌طور معمول، فاصله بین کلمات و علائم دستوری و نقطه‌گذاری به‌عنوان مرز کلمات در نظر گرفته می‌شود. روش‌های تعیین حد و حدود کلمات در نمایه‌سازی خودکار، بر اساس نوع برنامه و میزان پیشرفتگی آنها متفاوت است.

انواع نمایه‌سازی خودکار

کلولند[13] اشاره می‌کند که اصولاً دو نوع نمایه‌سازی خودکار وجود دارد. نمایه‌سازی انتسابی[14] و نمایه‌سازی اشتقاقی[15]. بنا به نظر کلولند نمایه‌سازی خودکار "عبارت است از استخراج ماشینی کلمات از درخواست‌ها و مدارک و انتساب ماشینی اصطلاحات کنترل شده و استخراج ماشینی از متون تمام متن یا بخشی از یک متن (۱۱). در این تعریف، نمایه‌سازی اشتقاقی یک فرآیند ساده انطباق بدون هیچ‌گونه تحلیل موضوعی است که در آن تمام کلمات مدرک یا درخواست استخراج شده و تنها کلمات ممنوعه از این فهرست خارج می‌شوند. نمایه‌سازی انتسابی فرآیند پیچیده‌ای است که با استفاده از تحلیل‌های آماری، کلمات و اصطلاحات به مدرک منتسب می‌شوند.

لنکستر نیز نمایه‌سازی خودکار را به دو دسته تقسیم می‌کند. گرچه لنکستر و کلولند، هر دو، تقسیم‌بندی یکسانی را ارائه کرده‌اند ولی نامگذاری آنها از این دو گروه نمایه‌سازی خودکار متفاوت است. لنکستر نمایه‌سازی خودکار را به دو دسته استخراجی[16] و انتسابی[17] تقسیم می‌کند. در ادامه این مقاله، این دو روش مورد بررسی قرار می‌گیرند.

نمایه‌سازی استخراجی

ساده‌ترین روش نمایه‌سازی مقالات در پایگاه‌های اطلاعاتی، روش نمایه‌سازی استخراجی است که در آن واژه‌ها برای قرار گرفتن در نمایه، توسط رایانه از متن مقالات استخراج می‌شوند. در این روش عموماً بسامد تکرار واژه در هر رکورد یا مقاله تعیین شده و کلماتی که بسامد تکرار آنها زیاد است در متن نمایه قرار می‌گیرند. کلمات پُربسامد همراه با تعداد رخداد آنها در کل پایگاه در متن نمایه قرار می‌گیرند(۸). این روش همچنین می‌تواند ریشه کلمات را شمارش کرده و کلمات منفرد یا عبارات را تشخیص دهد. می‌توان رایانه را به شیوه‌ای برنامه‌ریزی کرد که علاوه بر بسامد تکرار کلمات، با توجه به معیارهایی چون موقعیت کلمات در متن و معیارهای متنی دیگر، کلمات مورد نیاز برای نمایه‌سازی را به‌صورت خودکار استخراج نماید.

انواع نمایه‌سازی استخراجی

۱. نمایه‌سازی با استفاده از فهرست کلمات ممنوعه.

در این روش از نمایه‌شدن کلمات پُربسامد مشخصی جلوگیری شده و به‌صورت خودکار از فهرست کلمات نمایه حذف می‌شوند. کلماتی مانند: حروف اضافه، افعال کمکی، افعال،

حروف تعریف و ... در یک فهرست قرار می‌گیرند. در هنگام نمایه‌سازی، رایانه تمام کلمات متن را استخراج می‌کند، سپس کلمات ممنوعه را حذف و بقیه کلمات را در یک نظام الفبایی مرتب می‌کند.

تاکنون برای تعیین فهرستی از کلمات ممنوعه تحقیقات فراوانی صورت گرفته است. فرانسیس و کوسرا[18] ده کلمه پُربسامد را تعیین کردند. آنها همچنین فهرست دیگری که از ۴۲۵ کلمه ممنوعه تشکیل شده بود را ارائه دادند. وان رایجزبرگن[19] نیز فهرستی ارائه داد که از ۱۵۰ کلمه ممنوعه تشکیل شده بود. فهرست رایجزبرگن در سال ۱۹۷۵ منتشر شد.

همانطور که اشاره شد در این روش معمولاً همه کلمات متن به جز کلماتی که در فهرست کلمات ممنوعه قرار دارند در نمایه قرار می‌گیرند. استفاده از کلمات ممنوعه علاوه بر جلوگیری از ریزش کاذب در بازیابی، ۲۰ تا۲۵ درصد از حجم نمایه نیز می‌کاهد(۱۲).

روش دیگری نیز برای تهیه فهرست کلمات ممنوعه وجود دارد. در این روش فهرستی از کلمات متن به همراه بسامد تکرار آنها تهیه می‌شود. پس از تعیین بسامد تمام کلمات متن، کلمات پُربسامد بررسی و مطالعه می‌شوند. آن گروه از کلمات پُربسامدی که اهمیت اطلاعاتی نداشته باشند به‌عنوان کلمات ممنوعه در نظر گرفته می‌شوند. فهرست کلمات ممنوعه‌ای که مؤسسه ملی استاندارد و فن‌آوری[20] برای مجله وال استریت جرنال تهیه کرده نیز با این روش استخراج شده است. فهرست معمول کلمات ممنوعه که معمولاً در نمایه‌سازی خودکار به‌کار می‌روند در جدول زیر آمده است.

جدول ۱. فهرست عمومی کلمات ممنوعه

a	at	for	it	the	will
an	be	from	of	this	with
and	been	have	on	to	
are	but	in	or	was	
as	by	is	that	which	

۲. نمایه‌سازی بسامدی[21]. در این روش از نمایه‌سازی خودکار، بسامد تکرار کلمات در هر رکورد یا مقاله مورد بررسی قرار می‌گیرند و براساس بسامد تکرار در فهرست کلمات نمایه قرار می‌گیرند. پیشینه نمایه‌سازی استخراجی مبتنی بر بسامد تکرار کلمات به سال ۱۹۵۰ بازمی‌گردد. در این روش نمایه‌سازی، دفعات حضور واژه در یک متن شمارش می‌شود. سپس، واژگان استخراجی با فهرست کلمات ممنوعه مقابله و کلمات ممنوعه از فهرست نمایه خارج می‌شوند. سپس واژه‌ها بر اساس بسامد تکرار خود در نمایه مرتب می‌شوند.

علی‌رغم استفاده از فهرست کلمات ممنوعه، یکی از بزرگترین معایب نمایه‌سازی بسامدی آن است که هنوز شاهد حضور کلماتی در نمایه هستیم که علی‌رغم بسامد تکرار بی‌فایده‌اند. بعلاوه، گاهی حضور کلماتی در متون خاص علی‌رغم کمی تکرار از اهمیت فراوانی برخوردار است(۲). مثلاً، بسامد تکرار کلماتی چون "کتابخانه" و "اطلاع‌رسانی" یا "اطلاعات" در متون کتابداری چندان تعیین‌کننده نیست. اما حضور کلماتی چون "عایق‌کاری" و "پنبه‌نسوز" یا "کف‌پوش" در متون کتابداری علی‌رغم کمی تکرار می‌تواند بااهمیت باشد.

برای رفع مشکل نمایه‌های بسامدی، برای بسامد تکرار واژه حدنصابی قرار می‌دهند. بنابراین، کلماتی که بیش از حد نصاب تعیین‌شده در پایگاه تکرار شده باشند به‌عنوان کلمات نمایه برگزیده می‌شوند. حدنصاب بسامد، باتوجه به نوع متنی که کلمه در آن واقع شده می‌تواند متغییر باشد.

۳. نمایه‌سازی استخراجی به روش پسوندیابی[22] **یا ریشه‌یابی**[23]. در بعضی از سیستم‌های نمایه‌سازی استخراجی، از پسوند یا ریشه کلمات استفاده می‌شود. در این روش ریشه یا پسوند کلمات جایگزین مجموعه‌ای از کلمات هم‌ریشه یا پسوند مشترک می‌شود. الگوریتم‌های ریشه‌یابی مختلفی چون الگوریتم‌های استاندارد یا الگوریتم‌های موضوعی خاص همچون الگوریتم‌های پزشکی نیز وجود دارند. الگوریتم S یا استاندارد، ساختار مفرد و جمع کلمات را در هم ادغام می‌کند. الگوریتم Lovins، فهرستی از ۲۶۰ پسوند

موجود، فهرست عظیمی از موارد استثناء و شماری از قواعد را در بردارد. همچنین الگوریتم پسوندیابی Porter، ۶۰ پسوند را شامل می‌شود(۱۲).

۴. وزن‌دهی اطلاعات. استفاده از نظام وزن‌دهی در نمایه‌سازی خودکار، عموماً برای مجموعه‌های تمام‌متن با حجم زیاد بسیار کاربرد داشته و مفید است. گرچه بسیاری از سیستم‌های نمایه‌سازی رایانه‌ای از نظام وزن‌دهی اصطلاحات و کلمات متن بهره نمی‌گیرند، اما به‌نظر می‌رسد که این روش برای همه سیستم‌هایی که بر اساس احتمالات و رتبه‌دهی به انتخاب کلمات نمایه اقدام می‌کنند مناسب باشد. در این روش، کلمات بر اساس محل قرار گرفتن خود در متن (مثلاً عنوان، چکیده و ...)امتیازدهی می‌شوند. حضور کلمات در بخش‌های مختلف رکورد، امتیازات متفاوتی دارد. معمولاً حضور کلمه در عنوان مدرک بیشترین امتیاز را به خود اختصاص می‌دهد. این روش به‌دلیل کارایی پایین آن در بین بهره‌گیران مقبولیت چندانی ندارد. کارایی پایین این نظام نمایه‌سازی به دلایل زیر است:

۱. روشی برای گزینش کلمات مهم از درخواست وجود ندارد. مثلاً اگر فردی موضوع "وزن‌دهی اصطلاحات در بازیابی اطلاعات" را به سیستم وارد کند ممکن است تنها یک مدرک که هر چهار واژه موجود در درخواست را دربر داشته باشد بازیابی نماید. یعنی این سیستم نمی‌تواند کلمات مهم درخواست (در اینجا "وزن‌دهی" و "اصطلاحات") را تشخیص داده و صرفاً مدارک مرتبط با آنها را بازیابی کند.

۲. روشی برای گزینش کلمات مهم مدرک وجود ندارد. با توجه به مثال قبل، در درخواست اطلاعاتی برای موضوع "وزن‌دهی اصطلاحات در بازیابی اطلاعات" باید مدرکی بازیابی شود که مثلاً تنها بسامد تکرار واژه "وزن‌دهی" در آن بیشتر نباشد بلکه بالاترین بسامد تکرار این اصطلاح در مقابله با بسامد سایر اصطلاحات درخواست در همان مدرک سنجیده شود(۱۲).

۳. بلند بودن یا کوتاه‌بودن متن را مورد توجه قرار نمی‌دهد. در یک مدرک طولانی بدیهی است که ممکن است بسامد تکرار یک واژه بیشتر از مدرکی باشد که متن آن کوتاه است. بنابراین ممکن است مدرکی بسیار مرتبط صرفاً به‌دلیل کوتاهی متن آن از رتبه کمتری برخوردار باشد. البته با بهره‌گیری از روش‌های آماری بسیار پیچیده در تنظیم نتایج جستجو می‌توان تا حدود بسیار زیادی بر این مشکلات فائق آمد.

همانگونه که پیش از این نیز اشاره شد در نظام‌های رایانه‌ای بیشتر از روش‌های نمایه‌سازی استخراجی استفاده می‌شود.یکی از عمده‌ترین مشکلات این نمایه‌ها به ویژه هنگام استفاده در پایگاه‌های اطلاعاتی، عدم بازیابی کلمات درخواست به‌دلیل نبودن آن کلمه در نمایه پایگاه اطلاعاتی است. دلیل این امر آن است که بهره‌گیران معمولاً یا همه کلمات مترادف با اصطلاح موجود در درخواست را وارد نکرده‌اند و یا از مترادفات آن بی‌خبرند. بنابراین، بسیاری از مدارک مرتبط از دست می‌روند. برای رفع این معضل معمولاً طراحان پایگاه‌های اطلاعاتی توانایی‌های نمایه‌ای را با توانایی نرم‌افزاری درهم می‌آمیزند. یکی از روش‌ها، امکان نمایش نمایه و انتخاب واژه درخواستی توسط خود بهره‌گیر است. روش دیگر، استفاده از نظام بازخورد مرتبط[۲۴] است. این روش به بهره‌گیران اجازه می‌دهد تا مدارک مرتبط کمی را برگزینند. سپس از سیستم می‌خواهند تا با توجه به این مدارک، مدارک مرتبط بیشتری را بازیابی نمایند(۴). امروزه این روش در اینترنت و پایگاه‌های اطلاعاتی تمام‌متن کاربرد فراوانی دارد.

نمایه‌سازی انتسابی

در این روش معمولاً رایانه برای نمایه‌سازی از اصطلاحنامه یا کنترل واژگان بهره می‌گیرد. اصطلاحنامه فهرستی از همه سرعنوان‌هایی است که ممکن است در نمایه‌سازی مورد استفاده قرار گیرند. درواقع در نمایه‌سازی انتسابی برای هر واژه منتسب "پرونده‌ای"[۲۵] از کلمات و عبارات مرتبطی که به‌نظر می‌رسد در مدارکی که نیروی انسانی آنها را نمایه کرده‌اند مکرر بکار رفته تهیه می‌شود. چون هر اصطلاح در کنترل واژگان دارای پرونده کلمات است، بنابراین می‌توان از برنامه‌ای رایانه‌ای برای انطباق عبارت‌های مهم در مدرک با این مجموعه پرونده‌ها استفاده کرد و درصورت انطباق واژه موجود در مدرک با واژه‌های موجود در

پرونده‌های کلمات، اصطلاحی که پرونده به آن مربوط است را منتسب کرد(۲). پرونده کلمات، کلماتی هستند که با یافتن آنها در متن مدرک، می‌توان اصطلاح مربوط به آن پرونده را به آن مدرک اختصاص داد. مثلاً برای اصطلاح Child birth، پرونده کلمات باید لغات birth, labor, labour, delivery, baby, born و Childbirth را دربر داشته باشد. علاوه بر پرونده کلمات، رایانه از "معیار گزینش"[۲۶] نیز بهره می‌گیرد. معیار گزینش دستورالعملی است که براساس آن مشخص می‌شود که با چه ترکیب و چه تعداد تکرار از هر واژه، باید اصطلاح مربوط را به مدرک منتسب کرد. مثلاً معیار گزینش باید تعیین کند که اگر کلمه child birth ده بار در مقاله‌ای تکرار شده باشد آیا باید آن را در نمایه قرار داد یا خیر؟ معیار گزینش مشخص می‌کند که چه هنگام و با حضور چه کلماتی و با چند بار تکرار می‌توان اصطلاح اصلی را در نمایه وارد کرد.

این روش گرچه در نمایه‌سازی خودکار پیشرفتی به‌شمار می‌آید اما در عمل چندان ساده نیست. اول اینکه معیارهای ماشینی باید بسیار پیشرفته باشند. در این روش انتساب اصطلاح به مدرک بر اساس بسامد تکرار اصطلاح یا اصطلاحات پرونده است. اگر اصطلاح اصلی با بسامد معینی تکرار شده باشد، گزینش آن برای قرار گرفتن در نمایه مناسب خواهد بود اما اگر بسامد کلمات پرونده بیشتر از اصطلاح اصلی باشد، آیا باز هم با همین اطمینان می‌توان آن را به مدرک اختصاص داد(۸).

یکی از برنامه‌های بسیار پیشرفته برای نمایه‌سازی خودکار از نوع انتسابی را مؤسسه Biosis برای پایگاه اطلاعاتی خود تهیه کرده است. در اینجا بر عناوین مقالات موجود در مجله‌ها تأکید شده است. کلمات موجود در عنوان مقالات با مجموعه‌ای نحوی از حدود پانزده هزار اصطلاح زیست‌شناسی مقابله و مقایسه می‌شوند. سپس، این مجموعه اصطلاح به نوبه خود با ۶۰۰ سرشناسه مفهومی (که موضوعاتی نسبتاً اعم هستند) تطبیق داده می‌شوند. درصورت انطباق، رایانه یکی از سرشناسه‌های مفهومی را به مدرک نسبت می‌دهد.

اگرچه در طول سی سال گذشته در زمینه نمایه‌سازی خودکار انتسابی شاهد پیشرفت‌های زیادی بوده‌ایم اما همه این تجربیات با تعدادی اصطلاحات محدود و در محیطی آزمایشگاهی به نتیجه رسیده و هنوز بدون مداخله انسان کاربرد آنها در محیط‌های بزرگتر یا پایگاه‌های اطلاعاتی عظیم به واقعیت نپیوسته است. درواقع، هنوز از نمایه‌سازی انتسابی جز در تولید نمایه‌های چاپی استفاده نشده است.

درکل باید گفت که بسیاری از سیستم‌های نمایه‌سازی خودکار واقعاً خودکار نیستند تا بتوانند جایگزین نیروی انسانی شوند بلکه در جهت کمک به نمایه‌سازان طراحی می‌شوند. شاید بهتر آن باشد که این برنامه‌ها را نیز "نمایه‌سازی به کمک رایانه" بنامیم.

تأثیر نمایه‌سازی و روش‌های خودکار بر متخصصین اطلاع‌رسانی

باپیشرفته‌تر شدن هر چه بیشتر برنامه‌های رایانه‌ای و ظهور و ارائه هر چه بیشتر اطلاعات در ساختار الکترونیکی، در نهایت نمایه‌سازی به شیوه سنتی کمتر صورت خواهد گرفت. این نقصان و کاهش با افزایش تعداد پایگاه‌های اطلاعاتی و ازدیاد پروژه‌های نمایه‌سازی و چکیده‌نویسی در کوتاه مدت جبران خواهد شد(۸). باید از نیروی انسانی برای انجام کارهای مهمی که از طریق مطالعه درباره شیوه استفاده و استناد متون حاصل می‌آید بهره گرفت.

نمایه‌سازان بیش از گذشته باید گزینشی عمل کرده و درباره کیفیت متون در دست نمایه‌سازی و محتوای موضوعی آنها تصمیم‌گیری نمایند. می‌توان از مهارت نمایه‌سازی برای توسعه نظام‌های رایانه‌ای و بررسی و آزمایش خروجی رایانه‌ها بهره گرفت. برای تهیه اصطلاحنامه و آموزش نویسندگان برای نگارش متن‌های رایانه‌ای به شیوه‌ای که قابل بازیابی باشند می‌توان نمایه‌سازان را به کمک طلبید. نمایه‌سازانی که در حوزه‌های موضوعی رایانه، ویرایش، کتابداری و اطلاع‌رسانی و اطلاعات کتابشناختی تجربه دارند بهتر می‌توانند از مزایای این فرصت‌های تازه استفاده نمایند.

نمایه‌سازان به متخصصین و سوادآموزان رایانه‌ای بدل خواهند شد و از این طریق خواهیم توانست شکاف‌های موجود در راه سامان‌دهی دانش را شناسایی کرده و به روشی کارآمد این شکاف‌ها را از میان برداریم. برای دستیابی به این

مهم باید دانش رایانه را فراگیریم. نه‌تنها باید ابزارهای مختلف نمایه‌سازی رایانه‌ای را فراگیریم بلکه باید روش سازماندهی و استفاده الکترونیکی آن را نیز بیاموزیم تا بتوانیم به‌خوبی نیازهای اطلاعاتی را درک کرده و دین خود را ادا نماییم.

نمایه‌سازی خودکار و اینترنت

در طول ۳۰ سال گذشته، نمایه‌سازی خودکار به‌عنوان پاسخی برای انفجار اطلاعات مطرح شده است. همانگونه که شاهد رشد هر روزه نشر الکترونیکی هستیم نمایه‌سازی و کاوش منابع تمام‌متن به استانداردی بالفعل در بسیاری از برنامه‌های بازیابی اطلاعات به‌ویژه در اینترنت بدل شده است. دو مبحث بسیار بحث برانگیزی که امروزه در حوزه فن‌آوری اطلاع‌رسانی جریان دارد - یعنی جستجوی وب و مدیریت دانش - در حوزه نمایه‌سازی تغییراتی را بوجود آورده است. امروزه با نمایه‌سازی تمام متن، بازیابی اطلاعات به بازیابی اصل مدارک ارتقاء یافته است.

اینترنت از جنبه‌های مختلفی همچون حجم منابع، ناهمگونی و دایره لغات، واحدهای کاوشی، راهبرد کاوش و تنوع رسانه‌های اطلاعاتی مثل ابرمتن، و چندرسانه‌ای، چهره بازیابی اطلاعات را تغییر داده است. حجم نمایه در بین موتورهای کاوش هنوز هم یکی از نکات مهم و اصلی رقابت در بین تولیدکنندگان موتورهای کاوش است گرچه امروزه تهیه بهترین نمایه و نه بزرگترین آن مورد توجه قرار می‌گیرد. این موضوع باعث شده است تا موتورهای کاوش در نمایه‌های خود به گزینش سایت‌ها و صفحات وب روی آورند و بسامد روزآمدسازی سایت‌ها را افزایش دهند.

ظهور اینترنت تحولات فراوانی در حوزه نمایه‌سازی پدید آورده است. اینترنت بر سرعت، دقت، هوشمندی، قدرت مشارکت، کاربرپسندی، جهانی‌بودن، جامعیت و چندزبانی نمایه‌ها در موتورهای کاوش افزوده است(۱). ظهور اینترنت از جنبه‌های مختلف بر نمایه‌سازی خودکار تأثیرگذار بوده است. این موارد عبارتند از:

حجم مجموعه

حجم اطلاعات موجود بر روی اینترنت و به‌طور خاص بر روی وب هر روزه در حال افزایش است. بدیهی است هر چه حجم مجموعه افزایش یابد نمایه‌سازی آن دشوارتر شده و نیازمند استفاده از روش‌های کامل‌تر و پیشرفته‌تری است.

تفاوت‌های دایره لغات

معمولاً در یک مجموعه یا پایگاه اطلاعاتی یکدستی - چه ازنظر ساختار و چه از نظر موضوعی - وجود دارد. وب برخلاف پایگاه‌های اطلاعاتی، تقریباً از هیچ تجانس موضوعی و زبانی برخوردار نیست. علاوه‌براین، بهره‌گیران از اینترنت نیز بسیار متنوع‌اند. در یک تحقیق مشخص شد که احتمال استفاده یکسان دو فرد از اصطلاحی واحد در اینترنت در حدود ۲۰ درصد است(۱). اگرچه برخی از تولیدکنندگان صفحات وب در تلاش بوده‌اند تا با ایجاد ابرداده‌ها که نوعی اطلاعات طبقه‌بندی شده درباره محل اطلاعات است، ایجاد نمایه‌های خودکار را تسهیل نمایند اما هنوز استاندارد واحدی برای ایجاد این ابرداده‌ها وجود ندارد و نه‌تنها همه تولیدکنندگان وب خود را ملزم به تبعیت از آن نمی‌دانند بلکه بسیاری از آنها از وجود چنین استانداردهایی بی‌خبرند.

راهبردهای کاوش

با ظهور و همه‌گیری اینترنت و توسعه کاوش‌های درون‌خطی از مداخله و واسطه‌گری کاوشگران متخصص در بازیابی اطلاعات کاسته شده است. از سوی دیگر، تنوع بهره‌گیران که هر یک درخواست‌های اطلاعاتی متنوعی دارند و فقدان آگاهی لازم درباره شیوه کاوش در اینترنت و حتی بی‌اطلاعی از "راهبرد کاوش"، تولید نمایه‌ای خودکار و منسجم را دشوار ساخته است.

رسانه‌های جدید ارتباطی در اینترنت

امکانات تازه‌ای که در وب پدید آمده به گونه‌ای است که با سیستم‌های سنتی و اولیه در پایگاه‌های اطلاعاتی کاملاً متفاوت است. ایجاد امکانات فرامتنی با توانایی‌های منحصر به فردی چون "اتصالات[۲۷]"، خدمات چندرسانه‌ای و چندزبانه بودن اطلاعات اینترنت، بیش از گذشته ایجاد نمایه‌های خودکار را با دشواری مواجه ساخته است.

یادداشت‌ها

1- Microfilm
2- Microfish
3- Access Point
4- Controlled Indexing
5- Free Indexing Languages
6- Natural Language Indexing
7- Defence Technology Information Center (DTIC)
8- Russian International Center for Scientific and Technical Information
9- Computer Aided Indexing
10- Automatic Indexing
11- H. P. Luhn
12- Single Terms
13- Cleve Land
14- Assigned Indexing
15- Derived Indexing
16- Extraction
17- Assignment
18- Francis , Kucera
19- Van Rigsbergen
20- National Institute of Standards and Technology (NTST)
21- Frequency indexing
22- Suffixing
23- Stemming
24- Relevance Feedback
25- Prodile
26 - Criteria for inclusion
27- Links

منابع

1. Jacqmin, Laurence. "**Automatic indexing: how websearching and knowledge management hypes have changed the technological state of the art**". in Online information 98, Proceedings of 22th international information meeting, Learned Information, Oxford, PP>103-107.

2. Lancaster, F. W. "**Indexing and abstracting in theory and practice**". Illinois : University of Illinois, 1991.

3. Buchanan, Brian. "**A glossary if indexing terms**". London: Clive Bingley, 1976.

4. Rowley, E. "**Abstracting and indexing**". 2nd ed. London: Clive Bingley, 1988.

۵. میرزاده، احمد. "*نمایه و نمایه‌سازی*". ***نشریه فنی مرکز مدارک علمی***، دوره دوم، شماره دوم و سوم، ۱۳۵۲. ص. ۱۶-۲۹.

6. Baxendale, Phyllis B. "**Autoindexing by automatic processes**". Special libraries, 1965. 56(10), PP. 715-719.

۷. یوسفی، احمد. "*اصول و روش‌های نمایه‌سازی رایانه‌ای*". ***فصلنامه کتاب***، دوره نهم، شماره ۲ (تابستان ۱۳۷۷).

8. Browne, Glenda. "**Automatic indexing and abstracting**". www.autoindexing\automatic indexing and abstracting.htm

9. Collison, Robert L. "**Indexing and abstracting**". London: Ernest Benn, 1972.

10. **Auto-indexing of Keywords.** www.autoindexing\autoindexing of keywords.htm

11. Zheng, Fen. "**An ideal model of automatic indexing for patent systems**".

12. "**Challenges in indexing electronic text and images** / edited by Raya Fidel ... [et al]. Medford: Learned Information, 1994.

زبان و اطلاع‌رسانی

تألیف: حسن شکوهیان

مقدمه

در طول تاریخ بشر شیوه‌های اطلاع‌رسانی همواره دستخوش تغییر و دگرگونی بوده و در هر زمان متناسب با شرایط و امکانات به گونه‌ای خاص مورداستفاده قرار می‌گرفته است. با این حال یک اصل مشترک در همه این روش‌ها وجود داشته و آن عبارت است از استفاده از نوعی زبان (به مفهوم عام کلمه). اصولاً هرگونه پیام و اطلاعاتی از طریق بکارگیری نوعی از زبان منتقل می‌شود؛ به سخن دیگر بین اطلاعات و زبان پیوندی محکم و ناگسستنی وجود دارد. بدین خاطر بحث و بررسی در باب مسئله زبان یکی از دل‌مشغولی‌های متخصصان اطلاع‌رسانی بوده است. نوشته زیر سعی دارد تا مقوله زبان و ابعاد گوناگون آن را بررسی نماید.

تعریف زبان

زبانشناسان معتقدند که زبان را از دو دیدگاه مختلف می‌توان تعریف کرد:

۱ـ تعریف کاربردی ۲ـ تعریف قراردادی

تعریف کاربردی زبان: در این تعریف زبان عبارت است از یک وسیله مشترک اجتماعی برای بیان عقاید. این تعریف را می‌توان تعریف کاربردی زبان نامید زیرا بیانگر کاربرد زبان است.(۱)

تعریف قراردادی زبان: این تعریف، زبان را عبارت از