نمایه‌سازی براساس احتمالات یا نمایه‌سازی توزینی[1]

هارولد بورکو[2]
ترجمه دکتر منصوره باقری[3]

چکیده

نمایه‌سازی توزینی به‌منظور کاهش اختلالات معنایی دربازیابی اطلاعات طراحی شده است . به این ترتیب که هنگام نمایه‌سازی برای هرکلیدواژه ، وزنی مناسب با میزان ربط آن با محتوای مدرک تعیین می‌شود. در نمایه‌سازی توزینی ارتباط محتمل مدارک بازیابی شده برای پاسخگویی به سؤال محاسبه و همچنین میزان ربط احتمالی هریک تعیین می شود. زمانی که نتیجه جستجو به جستجوگر ارائه می شود، مدارک بر اساس میزان ربط احتمالی موضوع مدارک با موضوع سؤال از زیاد به کم، مرتب می‌گردد.

کلیدواژه‌ها

نمایه‌سازی احتمالی، نمایه‌سازی توزینی، ربط، بازیابی اطلاعات، دقت بازیابی

تاریخچه

واژه نمایه‌سازی توزینی نخستین بار در آگوست ۱۹۵۸ توسط مارون[4] در بولتن داخلی اداره " نظام های اطلاعاتی شرکت تامپسون راموولدریج" [5] به‌کار رفت (۳). حدود یک سال بعد مارون و همکارانش بولتن داخلی دیگری تهیه کردند و در آن تجاربی که برای ارزیابی کارایی نمایه‌سازی توزینی به‌کار گرفته بودند تشریح کردند (۵). نتایج این تجربه‌ها در نشریه Journal of the Association for Computing Machinery در جولای ۱۹۶۰ انتشار یافت (۴: ۲۱۶-۲۴۰). این مقاله علاوه بر اینکه توجه دبیزشگران را به مفهوم

۳.دکترای کتابداری و اطلاع‌رسانی و عضو هیئت علمی دانشگاه الزهرا mbagheri@alzahra.ac.ir

5.Data Systems Office of the Thompson Ramo Wooldridge Corporation

1."Probabilistic or Weighted Indexing". ***Encyclopedia of Library and Information Science***, Vol24. , pp: 212-220.
2.Harold Borko
4.M. E. Maron

نمایه‌سازی توزینی معطوف کرد، انگیزه پژوهش‌های دیگری در زمینه توسعه کارایی نظام‌های بازیابی ماشینی اطلاعات را ایجاد کرد.

علاوه برآن، این مقاله برای تشریح نظریه اولیه و شیوه‌های نمایه‌سازی توزینی، مفهوم و طریقه محاسبه میزان آن را که معیاری‌ست برای اندازه‌گیری احتمال ربط میان سؤال مطرح شده و مدرک مورد بازیابی، تعریف کرد. از آن پس تحقیقات جاری درباره نمایه‌سازی توزینی گسترش یافت و منجر به تدوین قوانین و مدل‌های توزین گردید. این مدل‌ها برای تمیز دادن کلماتی که در مقاله بار معنایی اصلی مرتبط با موضوع مدرک را دارند و به همین دلیل اصطلاحات مناسبی برای نمایه‌سازی هستند، از لغاتی که مفهوم خاصی را ندارند، به‌کار می‌رود. نمایه‌سازی توزینی نه تنها به فرایند فکری نمایه‌سازی وسعت دید می‌دهد بلکه جزء اصلی بیشتر نظام‌های بازیابی ماشینی‌ست.

نمایه‌سازی معمولی و نمایه‌سازی توزینی

در سال‌های آخر دهه ۱۹۵۰ با به‌کارگیری رایانه برای پردازش حروف همانند ارقام، آشکار شد که رایانه می‌تواند ترجمه‌ای اجمالی از زبانی به زبان دیگر انجام دهد و هم‌چنین به دلیل مقایسه میان واژه‌های تعیین شده در مدرک با واژه‌های مشابه موجود در سؤال، می‌توان از آن در بازیابی مدارک استفاده کرد، ولی نتیجه به‌دست آمده همیشه مطلوب نیست. پژوهشگر علم اطلاع‌رسانی پیوسته در جستجوی فنون جدیدی‌ست که بتواند کارایی نظام‌های بازیابی ماشینی اطلاعات را گسترش دهد. نمایه‌سازی توزینی یکی از این فنون است.

در نمایه‌سازی معمولی نمایه‌ساز باید تصمیم بگیرد که آیا یک واژه خاص را باید به یک مدرک اختصاص بدهد یا نه و آیا این واژه بیان‌کننده مضمون و محتوای مدرک هست؟ این انتخاب فقط دو حالت دارد، یعنی واژه‌ای که اختصاص می‌یابد ممکن است بیان‌کننده مضمون متن باشد یا نباشد. در هر حال هیچ‌گاه تطابق و هم‌خوانی میان واژه‌های نمایه و اطلاعات محتوای مدرک کامل نیست. یک کلیدواژه ممکن است برای نشان دادن جنبه‌ای از مدرک باشد که آن جنبه لزوماً موضوع اصلی مدرک نیست. علاوه بر این، محتوا ممکن است با مفاهیم مختلفی بیان شود، و یا یک واژه به مفاهیم مختلفی دلالت کند. عدم تطابق میان مفهوم و واژه را اختلال

جدول ۱. راهنمای تخصیص وزن های متحمل

وزن	شرح	کاربرد
۸/۸ : ۱/۰۰۰	موضوع اصلی	هنگامی که واژه بسیار خاص است و موضوع اصلی مدرک را در بر می گیرد.
۷/۸ : ۰/۸۷۵	موضوع اصلی	هنگامی که واژه خاص است وقسمت اعظم مدرک را در بر می گیرد.
۶/۸ : ۰/۷۵۰	موضوع کلی و عمومی	هنگامی که واژه مفهوم گسترده‌ای دارد و موضوع کلی‌تر را در بردارد.
۵/۸ : ۰/۶۷۵	واژه‌های مهم دیگر	هنگامی که واژه برای نمایه‌سازی به‌کار می‌رود ولی نه در موضوع اصلی.
۴/۸ : ۰/۵۰۰	موضوع کم اهمیت تر	هنگامی که واژه به مدرک مربوط است ولی موضوع اصلی مدرک را نمی‌پوشاند.
۳/۸ : ۰/۳۷۵	موضوع فرعی	هنگامی که واژه بیانگرنتایج آزمایش‌ها، روش‌ها و از این قبیل است.
۲/۸ : ۰/۲۵۰	موضوعات دیگر	هنگامی که واژه به نحوی با مدرک ا رتباط دارد.
۱/۸ : ۰/۱۲۵	ارتباط تقریبی	هنگامی که احتمال می‌رود جستجوگر با این واژه مطلب را جستجو کند ولی لازم نیست که برای طبقه بندی موضوع مورد استفاده قرارگیرد.

معنایی می‌گویند. استفاده از تزاروس می‌تواند این اختلال را کاهش دهد، ولی کاملاً از میان نمی‌برد چرا که قدری نامطمئنی در واژه‌های نمایه و موضوعی که به آن رهنمون شده‌اند وجود دارد، زیرا اینکه متقاضی اطلاعات با همان واژه خاصی که مدرک ذخیره شده است موافق باشد و آن را واژه مناسب جستجو بداند، یک احتمال است.

نمایه‌سازی توزینی وجود این اختلال معنایی را درک می‌کند، سعی برآن دارد که کارایی بازیابی را افزایش دهد. این کار به کمک اختصاص کلیدواژه‌ها بر اساس بهترین محاسبه احتمالات انجام می‌شود. اگر یک کلیدواژه فقط قسمتی از محتوای یک مدرک را نشان دهد آن واژه انتخاب می‌شود ولی به آن وزن اندکی مانند ۰/۲ یا ۰/۳ می‌دهند. در حالی که در نمایه‌های معمولی نمایه‌ساز باید تصمیم بگیرد این واژه را انتخاب کند یا خیر. انتخاب آن به این معناست که این واژه به‌طور کامل نشان‌دهنده محتوای متن است یا اینکه آن واژه را اختصاص ندهد؛ هر دو تصمیم می‌تواند در مواردی منجر به اشکال در بعضی بازیابی‌ها شود، برای مثال بازیابی مدرکی که مناسب نیست یا بازیابی نشدن مدرکی که مناسب بوده است . به همین دلیل نمایه‌سازی براساس احتمالات به کارایی بازیابی خواهد افزود؛ زیرا نمایه‌ساز مجاز است اصطلاحاتی را که قادر به توصیف بخشی از محتوای موضوعی مدرک است با وزنی که به لغات می‌دهد، تعیین کند. درنتیجه وظیفه نمایه‌ساز ساده‌تر و مطمئناً منطقی‌تر خواهد بود. به عبارت خلاصه‌تر، نمایه‌سازی توزینی سازوکاری‌ست که برای واژه‌های توصیف‌کننده مدرک یا درخواست، بار اطلاعاتی قائل شده میزان آن را تعیین می‌کند. این بار اطلاعاتی یا وزن واژه به‌وسیله نمایه‌ساز به‌طور دستی محاسبه و تخمین زده می‌شود، W_{ij} ، یعنی اگر کسی بخواهد نوع اطلاعات موجود در مدرک D_i را درخواست کند از کلیدواژه تعیین شده I_j استفاده خواهد کرد. بدیهی‌ست همین روش برای تحلیل و توزین درخواست نیز به‌کار می‌رود، گرچه دلیل تعیین وزن کمی متفاوت است.

اطلاع‌رسان با تخمین میزان احتمال اینکه کلیدواژه تا چه حد متقاضی را به مدرک مورد نیازش راهنمایی خواهدکرد، کلیدواژه را وزن می‌کند و به مدرک اختصاص می‌دهد. راهنمای توزین واژه‌ها هنگام نمایه‌سازی مدارک تهیه شده است و در جدول ۱ نشان داده می‌شود.

برای مثال، W_{ij}=۰/۸ به این معناست که بعد از تحلیل مدرک، نمایه‌ساز تخمین می‌زند اصطلاح j-th، یک شاخص ویژه است که به‌خوبی موضوع اصلی مدرک i-th را بیان می‌کند. برعکس برای کلیدواژه‌ای که ربط کمی دارد وزن W_{ij} = ۰/۵ تعیین می‌شود. باید توجه داشت که توزین به‌طور دستی و براساس استدلال عقلی صورت می‌گیرد.

به این ترتیب نمایه‌سازی توزینی از نمایه‌سازی معمولی متفاوت می‌شود. در نمایه‌سازی معمولی، نمایه‌ساز یک حق تصمیم دارد: همه یا هیچ. ولی در نمایه‌سازی توزینی این امکان هست که هر واژه‌ای را انتخاب کرد و به آن وزنی از ۰/۱ تا ۱/۰ داد. هشت تقسیم فرعی که در جدول توزین وجود دارد، نمایه‌ساز را در انتخاب وزن واژه یاری می‌دهد (۵: ۶۳).

توزین ربط و تعیین عدد ربط

هدف نمایه‌سازی توزینی، گسترش کارایی نظام‌های ماشینی ذخیره و بازیابی اطلاعات از طریق بالا بردن سطح جامعیت و مانعیت[۶]، کاهش اختلال معنایی و کمک به میسرکردن درجه‌بندی مدارک براساس میزان ربط با اطلاعات درخواستی‌ست. برای رسیدن به این هدف، در آغاز به یک معیار کمی برای اندازه‌گیری ربط نیاز است که باآن بتوان تصمیم گرفت. برای مثال برای سؤال R مدرک D_1مناسب‌تر از مدرک D_2 است. مشکل یا مسئله اندازه‌گیری میزان تناسب مدارک مانند سنجش مقدار اطلاع موجود دریک پیام است. شانون[۷] در اثر خود درباره نظریه اطلاعات توانست مقدار اطلاعات یک پیام را با شرایط احتمالی اندازه‌گیری کند (۸). میزان ربط را به‌وسیله فرمول بایس[۸] نیز می‌توان تعیین کرد. این فرمول که از طریق اجرای طرح به گونه‌های مختلف و با استفاده از حساب ساده احتمالات به‌دست آمده است، چنین است (۴: ۲۲۱):

$$P(A.I_j\ D_i) = \frac{P(A,D_i) \cdot P(A.D_i I_j)}{P(A,I_j)}$$

۶. زمانی که مفهوم نمایه‌سازی احتمالی برای نخستین بار عرضه شد، هنوز نسبت جامعیت و مانعیت برای سنجش میزان کارایی بازیابی رایج نشده بود.

7.C. E. Shanon
8.Bayes

$P(A.I_j\ D_i)$ یعنی احتمال اینکه استفاده‌کننده از کتابخانه سؤالش را با کلیدواژه I_j مطرح کند و مدارک D_i پاسخ رضایت‌بخش برای درخواست او باشد. بنابراین مدرک D_i می‌تواند برای سؤال مطرح شده، مناسب محسوب شود.

در این فرمول موارد زیر در نظر گرفته شده‌اند:

A یک مراجعه است. در این مراجعه از کتابخانه اطلاعاتی درخواست شده است.

D_i احتمال به دست آوردن مدرک و مرتبط یافتن آن است.

I_j احتمال درخواست اطلاعات دریک موضوع خـاص با به‌کار گرفتن اصطلاح نمایه j-th است، یعنی I_j موارد به صورت‌های زیرمی توانند ترکیب شوند:

$P(A,I_j)$ احتمال اینکه مراجعه‌کننده هنگام درخواست اطلاعات از نظام کتابخانه سؤالش را با اصطلاح I_j مطرح کند. اصطلاح I_j براساس تواتر استفاده در کتابخانه تعیین و تثبیت شده است.

$P(A,D_i)$ احتمال مفروضی که هنگام درخواست اطلاعات از نظام کتابخانه، مدرک D_i بازیابی خواهد شد. ارزش آن از تعداد دفعات استفاده از مدرک D_i تقسیم بر جمع تعداد دفعات استفاده از مدرک تعیین می‌شود. احتمال مفروض با استفاده از آمار در نظام کتابخانه بنا شده است.

$P(A.D_iI_j)$ احتمال اینکه مراجعه‌کننده، اطلاعاتی از نوع محتوای مدرک D_i را با اصطلاح I_j درخواست نماید. ارزش این اصطلاح به‌وسیله وزنی که مدرک i-th با واژه j-th نمایه شده است، برآورد می‌شود.

$P(A.I_jD_i)$ احتمال این‌که مراجعه‌کننده اطلاعات مورد درخواستش را با اصطلاح I_j مطرح کند و مدرک D_i با آن ربط داشته باشد. این میزان ربط است و مساوی‌ست با احتمال مفروض بازیابی مدرک D_i هنگام درخواست اطلاع، ضرب در احتمال درخواست مراجعه‌کننده از طریق کلیدواژه I_j ، تقسیم بر تعداد دفعات استفاده از آن کلیدواژه در نظام کتابخانه.

همه ارقام ذکرشده را می‌توان برآورد کرد و مقدار ربط را محاسبه نمود.

تعیین اعتبار از راه آزمایش

برای داشتن روش محاسبه میزان ربط، لازم است اعتبار آن از طریق آزمایش تأیید شود تا معین گردد که آیا سازوکارهای نمایه‌سازی احتمالی کارایی بازیابی را بالا می‌برد یاخیر. مارون و همکارانش آزمایش‌هایی را انجام دادند که فرضیه اولیه آن این چنین بیان شده است : میزان ربط محاسبه شده، مقیاس ربط احتمالی مدارک به سؤال طرح شده است (۴: ۲۳۱-۲۴۰). در نتیجه، این فرضیه پایه را می‌توان مجموع سه فرضیه دانست:

H_1 : اگر مدرکی با درخواست ربط دارد، نمره بالای $w_i(R)$ به آن داده می‌شود.

H_2:اگر مدرکی نمره بالای $w_i(R)$ را دارد، به سؤال طرح شده ربط دارد.

H_3: روش نمایه‌سازی احتمالی نمره بالای $w_i(R)$را فقط به مدرک دلخواه می‌دهد یعنی مدرکی که با درخواست ربط

جدول ۲. مقایسه میانگین ربط محاسبه شده با درجه بندی دستی مدارک

دسته‌بندی مدارک	میانگین ربط واقعی	انحراف معیار
۱. بسیار مرتبط	۰/۱۱	۰/۰۴۳
۲. مرتبط	۰/۷۲	۰/۰۵۳
۳. تاحدودی مرتبط	۰/۵۴	۰/۰۴۳
۴. کمی مرتبط	۰/۴۰	۰/۱۱۰
۵. بی ربط	۰/۱۸	۰/۰۱۳

کامل داشته باشد.

بـــرای امتحـان این فـرضیه، یک کتـــابخـــانه آزمایشی بـــا ۱۱۰ مقاله انتخـــابی در زمینـــه فیـــزیک از مجله News Letter Science تأسیس شد. ابتدا مقاله‌ها را با یک روش متداول نمـــایه‌سازی کردند. کلیـــدواژه‌های کنتـــرل نشده در ۴۷ گروه از اصطلاحات دسته بندی شد. سپس مقاله‌ها با استفاده ازوزن‌های محتمل که به این ۴۷ اصطلاح داده شد، دوباره نمایه‌سازی شد. چهل سؤال نسبتاً گسترده، با مراجعه به گروهی از مدارک که تصادفاً انتخاب شده بود، طرح شد. وزن سؤال‌ها نیز تعیین گردید.

هر سؤال در کتابخانه آزمایشی، جستجو شد. همه مدارک مطلوب، بازیابی و میزان ربط آنها اندازه‌گیری شد. برحسب میزان ربط سؤال با مدارک از ۰ تا ۱ از بالا به پایین نمره دادند. علاوه براین چهار نفر مدارک بازیابی شده را مطالعه کردند و به مدرک خیلی مرتبط نمره ۱ (یک)، مرتبط نمره ۲ (دو)، تاحدودی مرتبط نمره ۳ (سه)، کم ارتباط نمره ۴ (چهار) و به مدرک بی‌ربط نمره ۵ (پنج) دادند.

نتایج آزمایش در جدول ۲ خلاصه شده است. یافته‌ها به‌وضوح نشان می‌دهد که طبق محاسبه، میزان ربط در گروه ۱ دارای بالاترین نمره است (بسیار مرتبط). هرچه میزان ربط کمتر شود، امتیاز هم کاهش پیدا می‌کند و پایین‌ترین نمره برای گروه ۵ (یعنی بی‌ربط) است. این نتایج، فرضیه مورد آزمایش را تأیید می‌کند و شواهد آماری ثابت می‌کند که نمایه‌سازی احتمالی و درجه‌بندی مدارک به ترتیب میزان ربط، کیفیت بازیابی را افزایش می‌دهد (۵: ۸۱).

نمایه‌سازی توزینی در نظام‌های ماشینی

آزمایش نهایی استفاده از نمایه‌سازی توزینی و محاسبه میزان ربط به شکل دستی انجام شد. به این ترتیب که مدارک دستی نمایه‌سازی شدند و وزن محتمل آنها به شکل دستی اندازه‌گیری شد. باوجود این ادعا که نمایه‌سازی توزینی مشکل‌تر از نمایه‌سازی معمولی نیست، هیچ نظام عملیاتی در مقیاس وسیع از نمایه‌سازی توزینی استفاده نمی‌کند. به هر حال این سازوکار در نظام‌های ماشینی تجربه شده است و به‌خصوص در نظام اسمارت[9] (۶) با نتایج عالی همراه بوده است.

اساس نمایه‌سازی توزینی استفاده از اصطلاحات توزین شده برای مشخص‌کردن موضوع یک مدرک یا یک درخواست است. علاوه براین نمایه‌سازی توزینی متضمن این نکته است که میزان ربط احتمالی مدرک با درخواست، قابل محاسبه است و حاصل جستجو را می‌توان در یک نظم اولویت‌بندی شده به‌جای نظم اتفاقی به جوینده عرضه داشت. این سه معیار؛ استفاده از اصطلاحات نمایه توزین شده، تشخیص ربط مدرک، و تنظیم حاصل جستجو (برون‌داد) در نظام اسمارت منظور شد و نتیجه آن افزایش کارایی بازیابی بود. برای انجام این کار گام‌های زیر برداشته می‌شود:

جدول ۳.مقایسه اسمارت - مدلارز نشانگر اهمیت نمایه سازی توزینی و رتبه بندی مدارک

روش آزمایش	جامعیت	مانعیت
مدلارز (کلیدواژه‌های اختصاص یافته نمایه‌سازی آموزش دیده)	۰/۳۱۱۷	۰/۶۱۱۰
اسمارت (استفاده از ریشه لغات توزین نشده)	۰/۱۸۱۴ (۴۲٪ ـ)	۰/۴۱۴۱ (۳۲٪ ـ)
اسمارت (نمایه‌سازی توزینی نمایش مدارک بر اساس میزان احتمال مرتبط بودن آن)	۰/۲۶۲۲ (۱۶٪ ـ)	۰/۴۹۰۱ (۱۹٪ ـ)

9.Smart

کارایی یک نظام بازیابی اطلاعات عمدتاً به شیوه مورد استفاده در نمایه‌سازی مدارک ذخیره شده مربوط می‌شود. هنگام جستجو، واژه‌های موجود در سؤال با کلیدواژه‌های مدرک مطابقت داده می‌شود و درصورت تطابق، مدرک بازیابی می‌شود

۱. ابتدا کلمات موجود در چکیده مدارک و درخواست‌ها مشخص می‌شود، لغات سیاهه بازدارنده از فهرست لغات متن خارج می‌شود، پسوندها از آخر کلمات برداشته می‌شود تا لغات هم‌ریشه را بتوان باهم جور کرد.

۲. وزن لغات متن براساس بسامد ظهور ریشه لغت در چکیده مدارک یا ساختمان درخواست‌ها تعیین می‌شود. این نمایه‌سازی به‌طور خودکار انجام می‌شود. توجه داشته باشید که معیار تعیین وزن برمبنای میزان شمول اصطلاح برموضوع مدرک با تعداد دفعاتی که در چکیده ظاهرمی‌شود، تغییر می‌کند.

۳. بردارهای حاصل از توزین ریشه لغات مدارک و سؤال‌ها باهم مقایسه می‌شود و ضریب همبستگی برای هر گروه سؤال و مدرک که روی بردارهای برابر انعکاس مشابه دارند، محاسبه می‌شود. در نتیجه میزان ربط حساب شده است.

۴. مشخصات مدارک به ترتیب نزولی ضریب همبستگی به مراجعه‌کننده ارائه می‌گردد. مدارک بازیابی شده برحسب ارتباط محتمل با مورد درخواست تنظیم می‌شود (۷: ۲۰).

استفاده از نمایه احتمالی ماشینی و درجه‌بندی مدارک، جامعیت و مانعیت کاررا در نظام اسمارت ارتقاء بخشیده است. نتایج در جدول ۳ خلاصه شده است. در عملیات واقعی با افزودن اصطلاحنامه تهیه شده به‌وسیله ماشین و بازخورد استفاده‌کننده، اصلاحات بیشتری در کار بازیابی حاصل شد. نمایه‌سازی احتمالی، در همه جنبه‌های اصلی، ارزش خود را در نظام‌های ماشینی ذخیره و بازیابی مدارک اثبات کرده است.

مدل‌های محتمل نمایه‌سازی ماشینی

در نمایه‌سازی توزینی به‌وسیله بوکشتاین[10]، سوونسون[11]، (۱) و هارتر[12] (۲) پیشرفت‌هایی از جهات مختلف ایجاد شد. در مفهوم اصلی، "نمایه‌سازی احتمالی" روشی دستی‌ست که در آن وزن اصطلاحات نمایه بر مبنای اینکه چقدر در متن مدرک به آن پرداخته شده باشد، تعیین می‌گردد. کاربرد جدید "نمایه‌سازی احتمالی"و"مدل‌های محتمل نمایه‌سازی" برای توضیح فرایند آماری یعنی، احتمالی برای مشخص‌کردن رتبه کلمات که محتوای مدارک را بیان می‌کند، تشخیص این کلمات از کلماتی که برای انتقال مفهوم مدارک رسا نیستند. نظر اساسی که به‌وسیله بوکشتاین، سوانسون و هارتر طرح شد این است که لغات بی‌محتوا که بار اطلاعاتی چندانی ندارد مثل : بررسی، تحقیق، گزارش، و نظایر آنها به‌طور اتفاقی در گروهی از مدارک تکرار می‌شود، در حالی که لغات یا مضامین تخصصی مثل : لیزر، بافت‌شناسی، و موارد مشابه در تعداد نسبتاً کمی از مدارک جمع می‌شود و به‌طور تصادفی در مدارک ظاهر نمی‌شوند. مدل محتمل طرحی‌ست که در آن توزیع لغات مدارک طبق یک روند متعادل بیان شده است (۱؛ ۲).

این جدول از سالتن اقتباس شده است (۷: ۲۱) و برای اولین بار در مقاله « مقایسه نوینی بین نمایه‌سازی معمولی (مدلارز) و پردازش متن (اسمارت)» در مجله جیسیس، دوره بیست و سوم، ۲ (مارس-آپریل ۱۹۷۲)[13] منتشر شد.

با آزمایش این مدل روی مجموعه ۶۵۰ چکیده مشخص شد که بیشتر لغات بی‌محتوا به‌طور اتفاقی پخش‌شده است و برای نمایه‌سازی مفید نیست. بنابراین نشان داده می‌شود که اصطلاحات نمایه کارآمد را می‌توان از اصطلاحات غیرنمایه‌ای تمیز داد. مدل نمایه‌سازی احتمالی یک مبنای آماری را برای انتخاب اصطلاحات مناسب نمایه به‌وسیله ماشین از چکیده مدارک فراهم می‌کند. این مدل را می‌توان

10.Bookstein
11.Swanson
12.Harter
13.Jasis, Vol. 23, No. 2 (Mar.-Apr. 1972)

در نظام‌های بازیابی کاملاً ماشینی به‌کار برد.

خلاصه

کارایی یک نظام بازیابی اطلاعات عمدتاً به شیوه مورد استفاده در نمایه‌سازی مدارک ذخیره شده مربوط می‌شود. هنگام جستجو، واژه‌های موجود در سؤال با کلیدواژه‌های مدرک مطابقت داده می‌شود و درصورت تطابق، مدرک بازیابی می‌شود. در مباحث نظری، همه مدارک بازیابی شده باید با نیاز سؤال‌کننده مرتبط باشد ولی به دلایلی در عمل این طور نیست. ممکن است مدارک و سؤال‌ها به‌خوبی تجزیه و تحلیل نشده باشند یعنی در انتخاب کلیدواژه‌ها اشتباه شده باشد. گرچه از این نوع اشتباهات رخ می‌دهد ولی قسمت عمده اشتباهات بازیابی این نیست. مسئله مهم‌تر و جدی‌تر اختلالات معنایی‌ست که به دلیل عدم ارتباط دقیق میان مفهوم شرح داده شده با کلیدواژه مجاز به‌وجود می‌آید. این کلیدواژه ممکن است عام‌تر یا خاص‌تر از مفهوم مورد نظر باشد و یا ممکن است فقط جنبه‌ای از محتوای مدرک را در بر بگیرد که موضوع اصلی نیست. نمایه‌سازی توزینی فنی‌ست که برای کاستن اختلالات معنایی طراحی شده است. به این ترتیب که برای اصطلاحات نمایه، توزینی مناسب با میزان ربط کلمات با موضوع محتوای مدرک تعیین می‌کند. به علاوه نمایه‌سازی توزینی مستلزم آن است که ارتباط محتمل همه مدارک بازیابی شده برای سؤال، محاسبه شود و احتمال میزان ربط هریک را تعیین نماید. زمانی که نتیجه جستجو به سؤال‌کننده ارائه می‌شود، مدارک براساس میزان ربط احتمالی موضوع مدارک با موضوع سؤال از زیاد به کم، مرتب می‌شود.

نتیجه تجربه اصلی مارون یک مدل نمایه‌سازی توزینی دستی‌ست که ربط مدارک را با موضوع محاسبه می‌کند. نتیجه تجربه نشان می‌دهد که نمایه‌سازی توزینی می‌تواند کارایی بازیابی را افزایش دهد.

سالتن نشان داده است که در نظام بازیابی ماشینی، وزن نمایه‌سازی می‌تواند به‌طور خودکار براساس حضور کلمات در چکیده مدارک مشخص شود و این وزن‌ها همانند وزن‌های نمایه‌سازی احتمالی در نظام‌های دستی می‌تواند تفسیر و مورد استفاده قرار گیرد. همچنین نشان داده است که نمایه‌سازی توزینی کارایی بازیابی را افزایش می‌دهد.

هارتر، بوکشتاین و سوآنسن مدل‌هایی را شرح داده‌اند که به کمک آنها می‌توان از متن مدرک کلماتی را که بار اطلاعاتی خوبی دارند و قادرند کلیدواژه‌های مناسبی برای نمایه باشند، از لغات کم‌بارتر متمایز نمود. نتایج به‌دست آمده از مجموعه این مطالعات نشان می‌دهد که نمایه‌سازی توزینی کارایی بازیابی را توسعه می‌دهد و می‌تواند به‌طور خودکار به‌وسیله ماشین و به‌عنوان بخشی از پردازش داده‌ها انجام گیرد. این روش که هم ساده و هم مؤثر است، در نظام‌های ماشینی ذخیره و بازیابی اطلاعات مقبولیت بیشتری خواهد یافت.

مسئله مهم‌تر و جدی‌تر اختلالات معنایی‌ست که به دلیل عدم ارتباط دقیق میان مفهوم شرح داده شده با کلیدواژه مجاز به‌وجود می‌آید. این کلیدواژه ممکن است عام‌تر یا خاص‌تر از مفهوم مورد نظر باشد و یا ممکن است فقط جنبه‌ای از محتوای مدرک را در بر بگیرد که موضوع اصلی نیست. نمایه‌سازی توزینی فنی‌ست که برای کاستن اختلالات معنایی طراحی شده است

منابع

1. Bookstein, A.; Swanson, D. R. “Probabilistic Models for Automatic Indexing”. ***JASIS***, Vol. 25, No. 5 (Oct. 1944): 312-318.

2. Harter, S. P. “A Probabilistic Approach to Automatic Keyword Indexing”. ***JASIS***,

Vol. 26, No. 4 (Jul.-Aug. 1975): 197-206.

3. Maron, M. E. ***Probabilistic Indexing: A Statistical Approach to the Library Problem***. Los Angeles: Thompson Ramo Wooldridge, Inc., 1958.

4. Maron, M. E.; Kuhns, J. L. "On Relevance, Probabilistic Indexing and Information Retrieval". ***JACM***, Vol. 7, No. 3 (Jul. 1960): 216-240.

5. Maron, M. E.; Kuhns, J. L.; Ray, L. C. ***Probabilistic Indexing: A Statistical Technique for Document Identification and Retrieval***, Thompson Ramo Wooldridge, Inc., Los Angeles, Calif., Technical Memorandum, No. 3, June 1959.

6. Salton, G. ed. ***The SMART Retrieval System***. Englewood Cliffs: Prentice-Hall, 1971.

7. Ibid. ***Dynamic Information and Library Processing***. Englewood Cliffs: Prentice-Hall, 1975.

8. Shannon, C. E.; Weaver, W. ***The Mathematical Theory of Communication***. Urbana: University of Illinois Press, 1949.