



نمایه‌سازی براساس احتمالات یا نمایه‌سازی توزینی^۱

هارولد بورکو^۲
ترجمه دکتر منصوره باقری^۳

چکیده

نمایه‌سازی توزینی به منظور کاهش اختلالات معنایی در بازیابی اطلاعات طراحی شده است. به این ترتیب که هنگام نمایه‌سازی برای هر کلیدواژه، 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 هر یک تعیین می‌شود. زمانی که نتیجه جستجو به جستجوگر ارائه می‌شود، 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

کردند و در آن تجاری که برای ارزیابی کارایی نمایه‌سازی توزینی به کار گرفته بودند تشریح کردند (۵). نتایج این تجربه‌ها در نشریه *of Computing Machinery Association* در جولای ۱۹۶۰ انتشار یافت (۴: ۲۱۶-۲۴۰). این مقاله علاوه بر اینکه توجه دبیرشگران را به مفهوم

واژه نمایه‌سازی توزینی نخستین بار در آگوست ۱۹۵۸ توسط مارون^۴ در بولتن داخلی اداره "نظام‌های اطلاعاتی شرکت تامپسون راموولدریج"^۵ به کار رفت (۳). حدود یک سال بعد مارون و همکارانش بولتن داخلی دیگری تهیه

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

5. Data Systems Office
of the Thompson Ramo
Wooldridge Corporation

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

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

علاوه بر آن، 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، به کار 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 مدرک 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 ۰.۱ راهنمای تخصیص وزن های متحمل

وزن	شرح	کاربرد
۸/۸ : ۱/۱۰۰۰	موضوع اصلی	هنگامی که واژه بسیار خاص است و موضوع اصلی مدرک را در بر می‌گیرد.
۷/۸ : ۰/۸۷۵	موضوع اصلی	هنگامی که واژه خاص است و قسمت اعظم مدرک را در بر می‌گیرد.
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 به کار 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 بیانگر نتایج آزمایش‌ها، روش‌ها و از این قبیل است.
۲/۸ : ۰/۲۵۰	موضوعات دیگر	هنگامی که واژه به نحوی با مدرک ارتباط 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 قرار گیرد.

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 بر آن 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 $0/2$ یا $0/3$ 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. در نتیجه وظیفه نمایه‌ساز ساده‌تر و مطمئن‌تر خواهد بود. به عبارت خلاصه‌تر، نمایه‌سازی توزینی سازوکاری است 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_j را درخواست کند از کلیدواژه تعیین شده I_j استفاده خواهد کرد. بدیهی‌ست همین روش برای تحلیل و توزین درخواست نیز به کار 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 خواهد کرد، 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، $W_{ij}=0/8$ به این معناست که بعد از تحلیل مدرک، نمایه‌ساز تخمین

می‌زند اصطلاح j -th، یک شاخص ویژه است که به خوبی موضوع اصلی مدرک i -th را بیان می‌کند. برعکس برای کلیدواژه‌ای که ربط کمی دارد وزن $W_{ij}=0/5$ 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 $0/1$ تا $1/0$ داد. هشت تقسیم‌فرعی 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 میسر کردن 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 بآن بتوان تصمیم گرفت. برای مثال برای سؤال R مدرک D_j مناسب‌تر از مدرک D_p است. مشکل یا مسئله اندازه‌گیری میزان تناسب مدارک مانند سنجش مقدار اطلاع موجود در یک 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):

$$P(A, I_j | D_j) = \frac{P(A, D_j) \cdot P(A, D_j | I_j)}{P(A, I_j)}$$

7.C. E. Shannon
8. Bayes

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

$P(A, I_j; D_i)$ یعنی احتمال اینکه استفاده کننده از کتابخانه سؤالش را با کلیدواژه I_j مطرح کند و مدارک D_i پاسخ رضایت بخش برای درخواست او باشد. بنابراین مدارک D_i می تواند برای سؤال مطرح شده، مناسب محسوب شود. در این فرمول موارد زیر در نظر گرفته شده اند: A یک مراجعه است. در این مراجعه از کتابخانه اطلاعاتی درخواست شده است.

D_i احتمال به دست آوردن مدارک و مرتبط یافتن آن است. I_j احتمال درخواست اطلاعات در یک موضوع خاص با به کار گرفتن اصطلاح نمایه j -th است، یعنی I_j موارد به صورت های زیر می توانند ترکیب شوند:

$P(A, I_j)$ احتمال اینکه مراجعه کننده 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، مدارک D_i بازیابی خواهد شد. ارزش آن از تعداد دفعات استفاده از مدارک D_i تقسیم بر جمع تعداد دفعات استفاده از مدارک تعیین می شود. احتمال مفروض با استفاده از آمار در نظام کتابخانه بنا شده است. $P(A, D_i, I_j)$ احتمال اینکه مراجعه کننده، اطلاعاتی از نوع محتوای مدارک D_i را با اصطلاح I_j درخواست نماید. ارزش این اصطلاح به وسیله وزنی که مدارک i -th با واژه j -th نمایه شده است، برآورد می شود.

$P(A, I_j; D_i)$ احتمال این که مراجعه کننده اطلاعات مورد درخواستش را با اصطلاح I_j مطرح کند و مدارک D_i با آن ربط داشته باشد. این میزان ربط است و مساوی است با احتمال مفروض بازیابی مدارک D_i هنگام درخواست اطلاع، ضرب در احتمال درخواست مراجعه کننده از طریق کلیدواژه I_j ، تقسیم بر تعداد دفعات استفاده از آن کلیدواژه در نظام کتابخانه. همه ارقام ذکر شده را می توان 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 نمایه سازی احتمالی کارایی بازیابی را بالا می برد یا خیر. مارون و همکارانش آزمایش هایی 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 می توان مجموع سه فرضیه دانست:

H_1 : اگر مدارکی با درخواست ربط دارد، نمره بالای $w_i(R)$ به آن داده می شود.

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

H_p : روش نمایه سازی احتمالی نمره بالای $w_i(R)$ را فقط به مدارک دلخواه می دهد یعنی مدارکی که با درخواست ربط

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

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

کامل داشته باشد.

برای امتحان این فرضیه، یک کتابخانه آزمایشی با ۱۱۰ مقاله انتخابی در زمینه فیزیک از مجله News Letter Science تأسیس شد. ابتدا مقاله‌ها را با یک روش متداول نمایه‌سازی کردند. کلیدواژه‌های کنترل نشده در ۴۷ گروه از اصطلاحات دسته بندی شد. سپس مقاله‌ها با استفاده از وزن‌های 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 بر این چهار نفر مدارک بازیابی شده را مطالعه کردند و به مدرک خیلی مرتبط نمره ۱ (یک)، مرتبط نمره ۲ (دو)، تا حدودی 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. با وجود 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^۹ (۶) با نتایج عالی همراه بوده است.

اساس نمایه‌سازی توزینی استفاده از اصطلاحات توزین شده برای مشخص کردن موضوع یک مدرک یا یک درخواست است. علاوه بر این 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 انفاقی به‌جوبنده عرضه داشت. این سه معیار؛ استفاده از اصطلاحات نمایه توزین شده، تشخیص ربط مدرک، و تنظیم حاصل جستجو (برون داد) در نظام اسمارت منظور شد و نتیجه آن افزایش کارایی بازیابی بود. برای انجام این کار گام‌های زیر برداشته می‌شود:

جدول ۳. مقایسه اسمارت - مدلاز نشانگر اهمیت نمایه‌سازی توزینی و رتبه بندی مدارک

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

کارایی یک نظام بازیابی اطلاعات عمدتاً به شیوه مورد استفاده در نمایه سازی مدارک ذخیره شده مربوط می شود. هنگام جستجو، واژه های موجود در سؤال با کلیدواژه های مدرک مطابقت داده می شود و در صورت تطابق، مدرک بازیابی می شود

نتایج در جدول ۳ خلاصه شده است. در عملیات واقعی با افزودن اصطلاحنامه تهیه شده به وسیله ماشین و بازخورد استفاده کننده، اصلاحات بیشتری در کار بازیابی حاصل شد. نمایه سازی احتمالی، در همه جنبه های اصلی، ارزش خود را در نظام های ماشینی ذخیره و بازیابی مدارک اثبات کرده است.

مدل های محتمل نمایه سازی ماشینی

در نمایه سازی توزینی به وسیله بوکشتاین^{۱۰}، سوانسون^{۱۱}، (۱) و هارتر^{۱۲} (۲) پیشرفت هایی از جهات مختلف ایجاد شد. در مفهوم اصلی، "نمایه سازی احتمالی" روشی دستی است که در آن وزن اصطلاحات نمایه بر مبنای اینکه چقدر در متن مدرک به آن پرداخته شده باشد، تعیین می گردد. کاربرد جدید "نمایه سازی احتمالی" و "مدل های محتمل نمایه سازی" برای توضیح فرایند آماری یعنی، احتمالی برای مشخص کردن رتبه کلمات که محتوای مدارک را بیان می کند، تشخیص این کلمات از کلماتی که برای انتقال مفهوم مدارک رسا نیستند. نظر اساسی که به وسیله بوکشتاین، سوانسون و هارتر طرح شد این است که لغات بی محتوا که بار اطلاعاتی چندانی ندارند مثل: بررسی، تحقیق، گزارش، و نظایر آنها به طور اتفاقی در گروهی از مدارک تکرار می شود، در حالی که لغات یا مضامین تخصصی مثل: لیزر، بافت شناسی، و موارد مشابه در تعداد نسبتاً کمی از مدارک جمع می شود و به طور تصادفی در مدارک ظاهر نمی شوند. مدل محتمل طرحی است 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 نمایه سازی معمولی (مدلارز) و پردازش متن (اسمارت)» در مجله جیسیس، دوره بیست و سوم، ۲ (مارس-آپریل ۱۹۷۲)^{۱۳} منتشر شد. با آزمایش این مدل روی مجموعه ۶۵۰ چکیده مشخص شد که بیشتر لغات بی محتوا به طور اتفاقی پخش شده است و برای نمایه سازی مفید نیست. بنابراین نشان داده می شود که اصطلاحات نمایه کارآمد را می توان از اصطلاحات غیرنمایه ای تمیز داد. مدل نمایه سازی احتمالی یک مبنای آماری را برای انتخاب اصطلاحات مناسب نمایه به وسیله ماشین از چکیده مدارک فراهم می کند. این مدل را می توان

۱. ابتدا کلمات موجود در چکیده مدارک و درخواست ها مشخص می شود، لغات سیاهه بازدارنده از فهرست لغات متن خارج می شود، پسوندها از آخر کلمات برداشته می شود تا لغات هم ریشه را بتوان باهم جور کرد.

۲. وزن لغات متن براساس بسامد ظهور ریشه لغت در چکیده مدارک یا ساختمان درخواست ها تعیین می شود. این نمایه سازی به طور خودکار انجام می شود. توجه داشته باشید که معیار تعیین وزن بر مبنای میزان شمول اصطلاح بر موضوع مدرک با تعداد دفعاتی که در چکیده ظاهر می شود، تغییر می کند.

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

۴. مشخصات مدارک به ترتیب نزولی ضریب همبستگی به مراجعه کننده ارائه می گردد. مدارک بازیابی شده برحسب ارتباط محتمل با مورد درخواست تنظیم می شود (۷: ۲۰).

استفاده از نمایه احتمالی ماشینی و درجه بندی مدارک، جامعیت و مانعیت کار را در نظام اسمارت ارتقاء بخشیده است.

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

در نظام‌های بازیابی کاملاً ماشینی به کار 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 در صورت 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 به خوبی 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 اختلافات معنایی است که به دلیل عدم ارتباط دقیق میان مفهوم شرح داده شده با کلیدواژه مجاز به وجود 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

فنی است که برای کاستن اختلافات 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 هر یک 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 دستی است 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 سوانسن 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.

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*,

مسئله مهم‌تر و جدی‌تر اختلافات معنایی است که به دلیل عدم ارتباط دقیق میان مفهوم شرح داده شده با کلیدواژه مجاز به وجود 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 فنی است که برای کاستن اختلافات معنایی طراحی شده است

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.

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.

