سامانهٔ هوشمند ارزش‌گذاری مدارک بازیابی شده در پایگاه‌های فارسی به وسیلهٔ سیستم خبره

محمدباقر دستغیب[1]
دکتر شهرام جعفری[2]

چکیده

هدف: هدف پژوهش حاضر، استفاده از سیستم خبره در افزایش کارآیی سیستم ارزش‌گذاری مقاله‌های بازیابی شده در موتورهای جستجوی پایگاه‌های فارسی است.

روش: روند کار به این صورت است که سیستم خبره ارزش مقاله‌های را محاسبه می‌کند و نتایج حاصل بهتر از سیستمهای مرسوم می‌شود. سیستم هوشمند پارامترهای انسانی را به صورت قوانین سیستم خبره برای محاسبهٔارزش مقاله‌ها به کار می‌برد و در نهایت کارآیی بهتری از سیستمهای غیرهوشمند به دست می‌آید، زیرا دسترسی به منابع در پایگاه‌های اطلاعاتی و سطح وب یکی از بزرگترین چالشهای سیستمهای اطلاعاتی و سیستمهای بازیابی اطلاعات است.

نتایج: اگر فهرست به صورت مناسب مرتّب نشده نباشد، با اینکه سیستم موفق به یافتن جواب پرسش کاربر گردیده، دسترسی برای کاربر مهیا نشده است. از این رو، استفادهٔ بهینه از ۱۰ پاسخ اول و جایگاه صفحه اول بسیار مهم است. نتیجهٔ سیستم هوشمند حاصل، کارآیی بهتری از سیستمهای سنتی غیرهوشمند دارد.

کلیدواژه‌ها: سیستم خبره، ارزش‌گذاری نتایج جستجو ، سیستم هوشمند ارزش‌گذاری، جستجو در سیستمهای بازیابی اطلاعات.

---

۱. دانشجوی دکترای مهندسی کامپیوتر– هوش مصنوعی– دانشگاه شیراز.dstghaib@srist.com

۲. استادیار بخش کامپیوتر، دانشگاه شیراز.

مقدمه

دسترسی به منابع و جستجوی اطلاعات مورد نیاز و یافتن مدارک مرتبط با پرسش[1]، یکی از دغدغه‌های موتورهای جستجو و پایگاه‌های برخط می‌باشد ، لذا برای آنکه یافتن مدارک مرتبط با پرسش، امکان‌پذیر باشد، ارزش‌گذاری[2] نتایج حاصل از جستجو بسیار مهم است.

فاکتوری که ارزش‌گذاری مدارک را پراهمیت می‌کند ، محل فیزیکی قرارگیری مدرک در فهرست مدارک ارائه شده به کاربر است. معمولاً کاربران صفحات ابتدایی و ۱۰ نتیجهٔ برتر (صفحه اول) را مشاهده می‌کنند. بنابراین، بسیار مهم است که نتایج مرتبط‌تر و دارای نرخ ارزش بالاتر در این فهرست ۱۰ تایی برتر قرار بگیرند. ولی محاسبهٔ ارزش مدارک به طور دقیق ممکن نیست.

اگر تعداد نتایج حاصل از پرسش کم باشد، به تابع ارزش‌گذاری مدارک بازیابی شده نیاز نیست و صرفاً با توجه به شباهت مقاله‌ها به پرسش، می‌توان آنها را مرتّب کرد. ولی اگر تعداد نتایج حاصل از پرسش زیاد باشد (در این صورت معمولاً پرسش دارای کلیدواژه‌هایی با کاربرد عمومی است) به کارگیری یک تابع ارزش‌گذاری برای مقاله‌ها بسیار ارزشمند است، تا مقاله‌های بازیابی شده بر اساس ارزش مقاله‌ها فهرست شود. این ارزش باید به طور نسبی محاسبه شود.

چالشهای مهم در ارزش‌گذاری مدارک بازیابی شده، عبارتند از:

- به دست آوردن توابعی که بتواند به طور خودکار مدارک را ارزش‌گذاری کند و آیا این توابع در زمینهٔ بازیابی اطلاعات به خوبی عمل می‌کند؟
- آیا این توابع برای زمانی که تعداد جوابها بسیار محدود است نیز صحیح عمل می‌کنند؟
- چگونه توابع را برای به دست آوردن K نتیجهٔ برتر در پایگاه داده‌های بزرگ اجرا کنیم؟

1. query.
2. Ranking.

برای ارزش‌گذاری مدارک در سیستمهای بازیابی اطلاعات، استفاده از فرکانس کلمات (TF)و فرکانس معکوس مدارک[1] مرسوم است. در این روش، منظور از فرکانس، بسامد کلمات، یا تعداد مشاهده کلمه در مدارک (مقاله‌ها) می‌باشد.

ازجمله روشهای دیگر برای ارزش‌گذاری مقاله‌ها، استفاده از بازخورد ربط[2] از کاربران و فیلترینگ مشارکتی مدارک بازیابی شده است. ولی می‌خواهیم روشی را پیشنهاد دهیم که به طور خودکار بتواند مقاله‌ها را ارزش‌گذاری کند. روش پیشنهادی در پایگاه مقاله‌های فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فنّاوری پیاده‌سازی گردید که نتایج آن با سیستم قبل مقایسه می‌شود.

استفاده از یک سیستم خبره[3] که بتواند عملیات ارزش‌گذاری را انجام دهد و سپس مدارک را آمادهٔ مشاهده کاربر نماید بسیار ارزشمند است، زیرا دانش فرد خبره و تجربه جمع کاربران را برای پیشبرد هدفها و رفع چالشها به کار می‌گیرد، زیرا محاسبهٔ دقیق ممکن نیست و استفاده از سیستم هوشمند به ابهام زدایی کمک می‌کند. در ادامه، ابتدا معماری سیستم پیشنهادی بررسی می‌شود و پس از آن فاکتورهای سیستم خبره مرور و در نهایت قوانین به کار رفته شرح داده شده و کارآیی سیستم پیشنهادی با سیستمهای مرسوم مقایسه گردیده است.

معماری سیستم

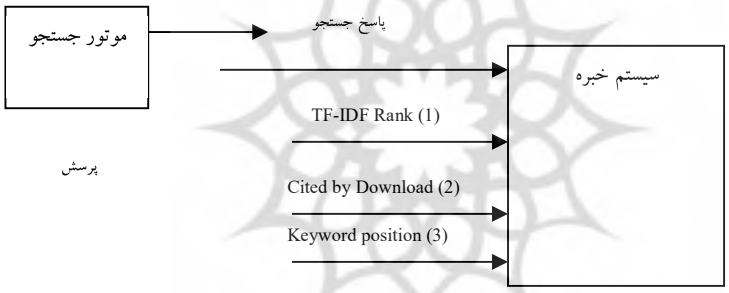
معماری سیستم بدین صورت است که یک موتور جستجو که در اینجا موتور جستجوی سامانهٔ مقاله‌های فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری است، اطلاعات مربوط به مقاله‌های فارسی را بر اساس پرسش کاربر جستجو می‌کند، سپس در این مرحله سیستم خبره پاسخ جستجو شده توسط موتور جستجو (پاسخ جستجو) را دریافت و نتایج حاصل از پرسش را، بر اساس قوانین داده شده به سیستم خبره

1. Inverse document frequency IDF.
2. Relevancy feedback.
3. Expert system.

ارزش‌گذاری می‌کند و ارزش نهایی هر مقاله را به طور تقریبی محاسبه و در نهایت جواب جستجو را بر اساس ارزش نهایی مقاله‌ها، فهرست و مرتّب می‌کند.

هدف سیستم خبره، ارتقای مقاله‌ها با ارزش بیشتر به بالای فهرست ارائه شده به کاربر است، با این فرض که کاربران معمولاً توجه وافری به ۱۰ نتیجهٔ برتر فهرست دارند. بنابراین، مقاله‌های شبیه‌تر و با ارزش‌تر برای پرسش کاربر باید در بالای این فهرست قرار گیرد.

وظیفهٔ سیستم خبره، جادادن مقاله‌های با ارزش‌تر در این فهرست ۱۰ تایی است. در ادامه، پارامترهای سیستم خبره و معماری آن شرح داده می‌شود. شکل ۱، معماری سیستم خبره را نشان می‌دهد.



شکل ۱. معماری سیستم

شکل ۱ معماری سیستم را نشان می‌دهد. نتیجه حاصل از موتور جستجو و سه پارامتر برای هر مقاله بازیابی شده ورودی سیستم خبره است. با کمک این پارامترها و به کارگیری مجموعه قوانین، ارزش نهایی مقاله‌ها محاسبه می‌شود و ده نتیجه برتر به دست می‌آید.

پارامترهای سیستم خبره عبارتند از:

۱. ارزش مقاله در موتور جستجو با استفاده از الگوریتم TF-IDF

۲. بازخورد ربط هر مقاله که با تعداد مراجعهٔ کاربر و یا دانلود برای هر مقاله محاسبه می‌شود.

۳. محل یافت شدن کلمات کلیدی (عنوان و یا کلیدواژه) و ارزش‌گذاری آنها.

قوانین سیستم خبره بر اساس این سه پارامتر نوشته شده‌اند و سیستم خبره بر اساس مقادیر این پارامترها تصمیم‌گیری می‌کند. پایگاه قوانین به صورت فازی[1] است؛ بنابراین مقادیر دریافتی ابتدا به ارزشهای فازی تبدیل و بر اساس پارامترهای فازی، قوانین فازی به کار گرفته می‌شود. نتیجهٔ نهایی نیز با تبدیل مقادیر فازی به عددی بین صفر و یک، که ارزش نهایی مقاله است، محاسبه می‌شود.

پارامترهای سیستم خبره

۱. ارزش مقاله در موتور جستجو با استفاده از الگوریتم TF-IDF:

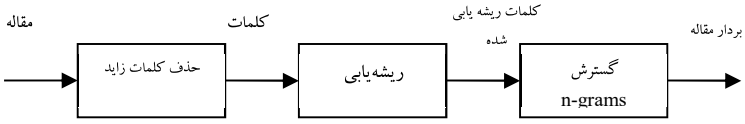
در موتور جستجوی مقاله‌های فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، از فضای برداری و الگوریتم TF-IDF برای ارزش‌گذاری مقاله‌ها استفاده شده است. بنابراین، لازم است یک سری عملیات پیش پردازش برای هر مقاله انجام شود، که در اینجا به اختصار به شرح این سلسله عملیات می‌پردازیم.

برای آنکه نویز مجموعهٔکلمات هر مقاله به حداقل برسد، ابتدا باید کلمات زاید (stop words) از فهرست کلمات حذف شود. این فرایند به وسیلهٔ فهرستی از کلمات زاید که در بازیابی اطلاعات بی‌تأثیر بوده و ارزش اطلاعاتی ندارد، انجام می‌شود. در این فهرست، کلماتی از قبیل حروف اضافه ربط و .... قرار دارد.

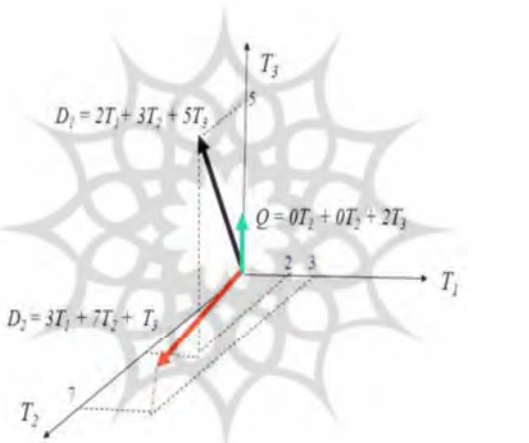
پس از حذف کلمات زاید، مجموعه‌ای از کلمات مقاله به دست می‌آید. حال باید این کلمات را ریشه‌یابی کنیم تا دقت جستجو حفظ شود. برای ریشه‌یابی کلمات می‌توان از الگوریتمهای شبیه پرتر (porter) استفاده کنیم. در نهایت، از الگوریتمهای گسترش کلمات برای گسترش کلمات ریشه‌یابی شده استفاده می‌شود. در سیستم پیشنهادی از الگوریتم (n-gram) استفاده می‌شود. (شکل ۲) با کمک این الگوریتم،کلمات بسط داده می‌شود تا بازیابی بهتر انجام شود. بنابراین، برای هر مقاله یک بردار به دست می‌آید،که طول آن به تعداد کلمات مجموعه پایگاه داده‌هاست. برای هر کلمه، تعداد تکرار در این مقاله و تمامی پایگاه محاسبه شده و به وسیلهٔ فرمول TF-

1. Fuzzy.

IDF برای هر کلمه یک عدد محاسبه می‌شود و در بردار نهایی قرار داده می‌شود. مجموعهٔ بردارهای مقاله‌ها، ماتریسی به نام ماتریس کلمه/مدرک را ایجاد می‌کند و پردازشها روی این ماتریس انجام می‌شود (شکل ۳).



شکل ۲. پردازش مقاله و تهیهٔ بردار مقاله



شکل ۳. بردارها در فضای سه بعدی

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

شکل ٤. ماتریس کلمه– سند

شکل ۳ نمودار بردارهای مقاله‌ها و پرسش را در فضای سه بُعدی نشان می‌دهد. همان‌طور که در شکل ٤ نشان داده شده است، برای مجموعه مقاله‌ها یک ماتریس به دست می‌آید. هرعدد w نشان دهندهٔ وزن هر کلمه در یک مدرک است. برای پرسش کاربر نیز همانند روش فوق یک بردار به دست می‌آید. حال زاویهٔ بین بردار پرسش و بردارهای مقاله‌ها در مجموعهٔ مقاله‌ها، نشان دهندهٔ میزان شباهت مقاله‌ها و پرسش است. هرچه این زاویه کمتر باشد، شباهت بیشتری وجود دارد. برای تبدیل عدد زاویه به عددی بین صفر و یک به عنوان معیار شباهت و یا ارزش مقاله‌ها، از کیسنوس زاویه بین بردارها استفاده می‌شود. این عدد به عنوان اولین پارامتر سیستم خبره در سیستم پیشنهادی استفاده شده است. در سامانهٔ موتور جستجوی مقاله‌های فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، از این پارامتر به عنوان ارزش مقاله‌ها استفاده می‌شود و مقاله‌ها بر اساس این عدد فهرست می‌شوند.

مشکل استفاده از این عدد به عنوان ارزش هر مقاله زمانی است که مقاله‌های بازیابی شده با پرسش کاربر زیاد باشد، بنابراین به دست آوردن نتیجهٔ مطلوب با نگاه کردن به صفحه اول بازیابی شده، موجب کاهش دقت و پارامتر بازخوانی[1] می‌شود.

بازخورد ربط

این پارامتر تعداد مراجعه به مقاله‌ها را در نظر می‌گیرد. به عبارتی، مقاله‌هایی که بیشتر مراجعه کننده داشته‌اند، برتر هستند. از این فاکتور در بسیاری از پایگاه‌ها، نظیر پایگاه Amazon نیز استفاده می‌شود. هرچه یک مقاله بیشتر دانلود شده باشد، نشان دهندهٔ ارزش بیشتر مقاله و توجه بیشتر کاربران به آن است. ولی باید توجه داشت که مقاله‌های قدیمی‌تر، نسبت به مقاله‌های جدید سابقه طولانی‌تری در پایگاه دارند و احتمالاً تعداد مراجعهٔ بیشتری خواهند داشت. برای رفع این مشکل، تعداد مراجعه به مقاله‌ها بر اساس تعداد سال حضور در پایگاه داده‌ها، طبق فرمول زیر نرمال‌سازی می‌شود:

1. Recall.

$$RF = \frac{No \quad of \quad downloads}{No \quad of \quad Years}$$

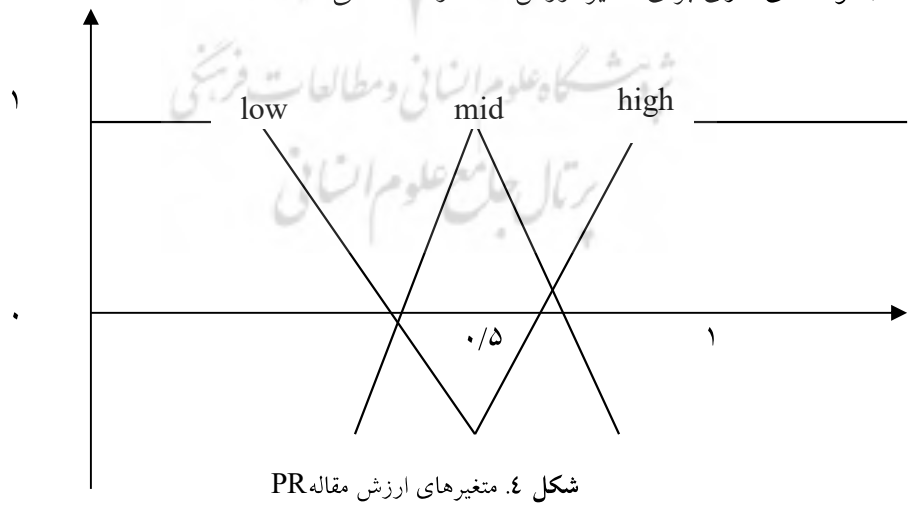
با توجه به فرمول فوق، میانگین تعداد مراجعه به هر مدرک با در نظر گرفتن تعداد سال حضور در پایگاه به عنوان پارامتر دوم محاسبه می‌شود.

محل یافت شدن کلمات کلیدی

محل یافتن کلمهٔ کلیدی مرتبط با پرسش برای محاسبهٔ ارزش مقاله، بسیار اهمیت دارد. برای محل قرار گرفتن کلمهٔ کلیدی، طبق قوانین سیستم خبره برای محل قرار گرفتن کلمهٔ کلیدی وزن در نظر گرفته می‌شود و به مقاله‌هایی که کلمهٔ کلیدی در عنوان آنها یافت شود، ارزش بالاتری داده می‌شود. این پارامتر در کنار دو پارامتر دیگر در تولید قوانین استفاده می‌شود.

قوانین سیستم خبره

قوانین به‌کار رفته در سیستم خبره، قوانین فازی است. بنابراین، متغیرهای قوانین، متغیرهای فازی هستند و در ابتدای کار باید مقادیر پارامترها به مقادیر فازی ترجمه شود. برای ترجمه مقادیر پارامترها ازمجموعه‌های فازی استفاده می‌شود. خروجی سیستم خبره نیز به صورت فازی است و در نهایت به عدد تبدیل می‌شود. شکل ٤ مجموعه‌های فازی برای متغیر ارزش مقاله را نشان می‌دهد.



شکل ٤. متغیرهای ارزش مقالهPR

به همین صورت، برای دو پارامتر دیگر نیز متغیرهای فازی تعریف می‌شود. سپس مجموعه قوانین سیستم خبره شکل داده می‌شود. مجموعهٔ قوانین سیستم خبره با متغیرهای فازی نوشته می‌شود. در ذیل، تعدادی از قوانین به کار رفته در سیستم خبره به عنوان مثال آورده شده است:

پارامتر ارزش مقاله‌ها در قوانین با PR ، پارامتر میانگین دانلود با DN و پارامتر محل کلیدواژه با KN نشان داده شده است. ارزش نهایی مقاله نیز با متغیر FR نشان داده شده است.

If PR is high and DN is high then FR is very high
If PR is high and DN is mid then FR is mid
If PR is high and DN is low then FR is mid
If PR is mid and DN is high then FR is high
If PR is mid and DN is mid then FR is mid
If PR is mid and DN is low then FR is low
If PR is low and DN is low then FR is very low
If KN is high and DN is high then FR is very high
...

با به کارگیری این قوانین، در نهایت متغیر FR ارزش نهایی مقاله را نشان می‌دهد. با تبدیل این متغیر به ارزش عددی بین صفر و یک، مقاله‌ها در یک فهرست مرتّب می‌شوند. برای مقایسهٔ نتایج با سیستم ارزش‌دهی موتور جستجوی سامانه مقاله‌های فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فنّاوری، ۱۰ نتیجهٔ برتر برای ۱۰۰ جستجو در نظر گرفته شده است.

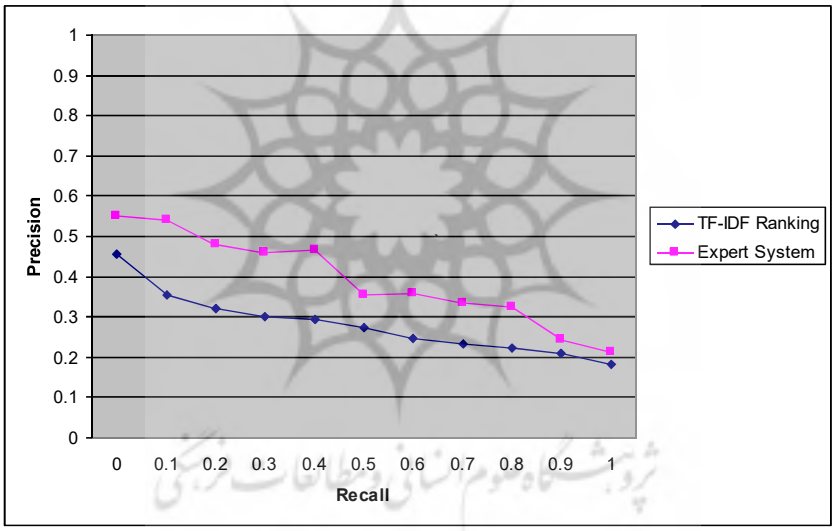
نتایج

برای مقایسهٔ نتایج، از دو معیار دقت و بازخوانی استفاده می‌شود. دقت عبارت است از نسبت تعداد مدارک مرتبط بر تعداد کل مدارک بازیابی شده بازخوانی نیز عبارت است از نسبت تعداد جوابهای بازیابی شده به کل تعداد جوابهایی که باید از پایگاه بازیابی می‌شده است. (جوابهای محتمل موجود در پایگاه)

این دو معیار برای صفحه اول با ۱۰ جواب در نظر گرفته و مقادیر به طور محلی محاسبه می‌شود و نمودار آن ترسیم می‌گردد. شکل ٦ نمودار دقت و بازخوانی را برای سیستم خبره و سیستم TF-IDF مقاله‌های فارسی (بدون سیستم خبره) نشان می‌دهد که

فقط برای صفحهٔ اول محاسبه شده است. با در نظر گرفتن صفحهٔ اول، مشخص است که سیستم خبره نتیجهٔ بهتری نسبت به سیستم مرسوم دارد.

باید در نظر داشت، پارامترها و مجموعه‌های فازی برای سیستم کنونی تنظیم شده است و در صورتی که مجموعه کاری تغییر کند، مجموعه متغیرهای فازی دوباره باید تنظیم گردد و همچنین قوانین فازی با توجه به فضای کاری مجموعه وزن‌دهی و بازنویسی شود. نقطه قوت سامانهٔ ارزش‌دهی هوشمند با استفاده از سیستم خبره، استفاده از چندیدن پارامتر در تصمیم‌گیری برای ارزش مکانی مقاله است که در مقایسه با ارزش‌دهی سیستم TF-IDF کیفیت بالاتری را ارائه می‌دهد. نقطه ضعف سیستم، زمان پاسخ بالاتر نسبت به سیستمهای غیر هوشمند است.



شکل ٦. نمودار مقایسهٔ سیستم خبره و سیستم ارزش‌دهی بر اساس TF-IDF

منابع

– دستغیب محمدباقر(۱۳۸۵). مروری بر نمایه‌سازی معانی پنهان: نظریه و کاربردها، فصلنامه کتابداری و اطلاع‌رسانی، شماره ۲۵ جلد ۷.

- ROSARIO B., "Latent Semantic Indexing: An overview", INFOSYS 240, spring 2000.
- Kowalski, G. "Information retrieval systems, Theory and Implementation", Kluwer Publisher. 1997.
- John Durkin (1994) , Expert systems Design and development, Macmillan Publishing Company.