مدل دو مرحله‌ای شکاف-گلچین برای نمایه‌سازی خودکار متون فارسی

محمد توکلی‌زاده راوری
استادیار گروه علم اطلاعات و دانش‌شناسی دانشگاه یزد
tavakoli@yazd.ac.ir
تاریخ دریافت: ۱۳۹۲/۸/۰۴؛ تاریخ پذیرش: ۱۳۹۳/۰۲/۲۰

چکیده

هدف: به علت خاص بودن برخی از مسائل زبانی، لازم است که مدل‌های بومی نمایه‌سازی خودکار را با توجه به ویژگی‌های هر زبان طراحی کرد. این مدل‌ها باید به‌گونه‌ای طراحی شود که جامعیت و مانعیت نمایه‌سازی مورد توجه باشد. هدف این مقاله معرفی و سنجش توانمندی مدل دو مرحله‌ای شکاف – گلچین برای نمایه‌سازی خودکار مقالات فارسی است. ابتدا الگوریتم کار به تفصیل توضیح داده می‌شود و سپس همخوانی نتایج حاصل از این الگوریتم با کلیدواژه‌های نویسنده سنجیده خواهد شد.

روش: مدل نمایه‌سازی خودکار فارسی به‌همراه توضیح مراحل و مسائل مرتبط با آن معرفی خواهد شد. ارزیابی مدل از طریق شاخص دربردارندگی انجام می‌شود که برای تعیین درصد همخوانی بین نمایه‌سازان مورد استفاده قرار می‌گیرد. برای این کار، میزان همخوانی اصطلاحات نمایه‌ای که از پیاده‌سازی الگوریتم این مدل حاصل شده‌اند، با کلیدواژه‌های نویسندگان مقالات بررسی می‌گردد.

یافته‌ها: یافته‌ها نشان داد که در ۹۰ درصد از موارد، اصطلاحی که این مدل در یک مقاله به‌عنوان پروزن‌ترین اصطلاح تشخیص داده است، مشابه اولین کلیدواژهٔ نویسنده آن مقاله است. در کل، بین نتایج این مدل و کلیدواژه‌های نویسندگان ۷۶ درصد همخوانی وجود داشت که در مقایسه با کارهای قبلی، قابل قبول به نظر می‌رسد.

اصالت/ارزش: ارزش اولیهٔ این کار پرداختن به نمایه‌سازی خودکار با توجه به ویژگی‌های زبان فارسی است. برای پیاده‌سازی مدل ارائه شده، فرض بر استفاده از زبان عبارات الگودار است که توسط بسیاری از زبان‌های برنامه‌نویسی پشتیبانی می‌شود و نیاز به نصب و استفاده از جدول‌های بانک اطلاعاتی را برای پردازش متن کاهش می‌دهد. همچنین، مشکل تعیین آستانهٔ بالایی اصطلاحات اصلی را حل می‌کند. علاوه بر آن، با الگوریتمی خاص، حد پایینی را نیز تعیین می‌کند؛ به‌گونه‌ای که دیگر تعداد اصطلاحات گلچین شده به طول متن بستگی ندارد. این امکان، جامعیت و مانعیت نمایه‌سازی را تضمین می‌کند.

کلیدواژه‌ها: نمایه‌سازی خودکار، زبان فارسی، مدل شکاف – گلچین.

تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی
فصلنامهٔ علمی- پژوهشی
دورهٔ ۲۱، شمارهٔ ۱، پیاپی ۸۰، بهار ۱۳۹۴
ص ص ۱۳- ۴۰
نمایه‌شده در ISC, SID و MagIran

Research on Information Science and Public Libraries
The Quarterly Journal of Iran Public Libraries Foundation
Vol. 21, No.1, Successive No.80 Spring 2015
Pp. 13-40
Indexed in ISC, SID & MagIran

مقدمه

نمایه‌سازی از نگاه متخصصان در دو معنای عام و خاص به کار می‌رود. در معنای عام می‌توان نمایه‌سازی را فرایند تجزیه و تحلیل داده‌ها و اطلاعات کتابشناختی مدارک دانست. هدف از این کار، پردازش داده‌های کتابشناختی برای وارد کردنِ آن به یک سامانهٔ اطلاعاتی است. معمولاً، منظور از نمایه‌سازی، تجزیه و تحلیل موضوعی مدارک است که از معنای خاص آن حکایت دارد. «هدف از نمایه‌سازی موضوعی، تعیین اصطلاحات موضوعی در قالب‌هایی مانند توصیفگر، سرعنوان موضوعی، شمارهٔ بازیابی، کد طبقه‌بندی یا اصطلاح نمایه‌ای با هدف بازیابی اطلاعات است» (مائی[1]، ۲۰۰۵). این اصطلاحات موضوعی، سلسله‌ای از «واحدهای زبانی»[2] هستند که «محتوای یک مدرک»[3] را نشان می‌دهند (وو[4]، آو[5] و ژانگ[6]، ۲۰۰۸).

یکی از معانی خاص نمایه‌سازی آن است که از چه عامل یا واسطه‌ای برای نمایه‌سازی استفاده شده است. از این منظر، می‌توان روش‌های نمایه‌سازی را به سه دسته تقسیم کرد که در هر کدام، عامل نمایه‌سازی «انسان» یا «ماشین» و یا ترکیبی از این دو است. این سه دسته عبارتند از: «دستی»، «ماشینی» و «نیمه ماشینی». به جای نمایه‌سازی ماشینی و نیمه‌ماشینی، می‌توان از معادل نمایه‌سازی خودکار و نیمه خودکار استفاده کرد.

در فرایند تشخیص و ترجمهٔ مفاهیم یک متن، دو ویژگی را باید مد نظر قرار داد که عبارت است از جامعیت[7] و مانعیت[8] اصطلاحات. «جامعیت، به این اشاره دارد که اصطلاحات موضوعی تا چه میزان محتوای یک مدرک را تحت پوشش قرار می‌دهند» (جونز[9]، ۲۰۰۴). بنابراین، «نمایه‌سازی هنگامی جامع است که منجر به تولید تعداد زیادی اصطلاح شود تا همهٔ جنبه‌های موضوعی مدرک را منعکس کند» (راغوان و دیگران[10]، ۲۰۰۴). بنا به گفتهٔ سوئرگل[11] (۱۹۹۴)، متوسط تعداد موضوعاتی که به مدارک در یک پایگاه اطلاعاتی داده می‌شود، میزانی برای اندازه‌گیری جامعیت است. لذا در طراحی هر سامانهٔ نمایه‌سازی، سیاستگذاری یا طراحی یک مدل برای کنترل سطح مناسب جامعیت لازم است. باید توجه داشت که جامعیت ترکیبی از دو جزء «نقطه نظر» و «اهمیت» است. «جامعیتِ نقطه‌نظری، به این پرسش می‌پردازد که آیا اصطلاحات نمایه‌سازی‌شده، در بردارندهٔ چهریزه‌ها یا نقطه‌نظرهای

1. Mai
2. unithood
3. Termhood
4. Vu
5. Aw
6. Zhang
7. exhaustificity
8. specificity
9. Jones
10. Raghavan
11. Soergel

مفید برای بازیابی هستند یا خیر. میزان اطمینانی که می‌توان به این پرسش جواب «بله» داد، میزان جامعیت نقطه‌نظری اصطلاحات است. جامعیتِ اهمیت، به این پرسش توجه دارد که در قوانین وضع شده برای نمایه‌سازی، تا چه حد اصطلاحات داده شده به یک مدرک اهمیت دارند. برای یک نمایه‌ساز، این پرسش این‌گونه مطرح می‌شود: کدام مفاهیم مرتبط با مدرک مورد نظر، به اندازهٔ کافی مهم هستند تا وجود آنها بتواند [کیفیت] نمایه‌سازی را تضمین کند (مارون[1]، ۱۹۷۹)» و به عبارتی، آیا مفاهیم مهم مدرک استخراج شده است یا خیر.

لنکستر[2] (۱۹۹۱) معتقد است که باید دقیق‌ترین اصطلاحی را برگزید که به صورت تمام و کمال محتوای مدرک را تحت پوشش قرار می‌دهد. مثلاً اگر مدرکی به دو نوع گُل پرداخته است، به جای اصطلاح گل‌ها، نام هر یک از گل‌ها آورده شود و اگر چندین گل را تحت پوشش قرار می‌دهد، در این جا اصطلاح «گل‌ها» دقیق‌تر خواهد بود. منظور از مانعیت نمایه‌سازی، تعریفی است که از لنکستر نقل شد. بنا به پژوهش گیه اونگ[3] (۲۰۰۶)، ارتباط محکمی بین مانعیت اصطلاحات نمایه‌ای و مرتبط بودن مدارک بازیابی شده وجود دارد. لذا، اگر سطح مانعیت اصطلاحات نمایه‌ای پایین باشد، وجود آنها در مرحلهٔ بازیابی مفید نخواهد بود. جنوین[4] و فلوید[5] (۲۰۰۴) به جامعیت و مانعیت نمایه‌سازی از زاویهٔ بازیابی اطلاعات نگاه کرده‌اند و به جای اصطلاح جامعیت از «حساسیت[6]» استفاده نموده‌اند. آنها حساسیت را توانمندی جستجو در بازیابی مقالات مرتبط و مانعیت را توان جستجو در بازیابی نکردن مقالات نامرتبط می‌دانند. بر اساس پژوهش آنها که روی جستجو بر اساس سرعنوان‌های موضوعی «مِش» متمرکز بود، جامعیت اصطلاحات داده شده به مدارک توسط نمایه‌سازان انسانی، چیزی بین ۵ تا ۳۶ درصد و مانعیت آنها بین ۸۵ تا ۹۹ درصد بود.

همچنین، در نمایه‌سازی باید از «افزونگی» دوری جست. نباید یک مفهوم عام را به عنوان اصطلاح نمایه‌ای برگزید و هم‌زمان، موضوع اخص نیز به آن داد. مثلاً، اگر به مدرکی اصطلاح «مدیریت اطلاعات» داده شده است، دادن اصطلاح اخص «سامانه رایانه‌ای مدیریت اطلاعات» نوعی افزونگی است. بنابراین، باید از میان دو مفهوم عام و خاص یکی را برگزید که جامعیت و مانعیت را تضمین کند.

1. Maron
2. Lancaster
3. Giyeong
4. Jenuwine
5. Floyd
6. sensivity

برای طراحی یک سامانهٔ نمایه‌سازی، باید موارد بالا مدنظر قرار گیرد تا اصطلاحات استخراج شده جامع‌ترین و دقیق‌ترین موضوعات باشند چون در نمایه‌سازی، تشخیص اصطلاحات مناسب از اهمیت زیادی برخوردار است. تشخیص خودکار اصطلاحات به سه روش اصلی انجام می‌شود که شامل روش‌های الف) دستور زبانی، ب) بازخوردی و پ) آماری می‌شوند:

الف) روش دستور زبانی بر این پیش‌فرض استوار است که منظور یک مدرک در ساختار دستوری جملات آن نهفته است. با توجه به نقش و نوع ترکیبات اسمی در جمله می‌توان منظور یک مدرک را تشخیص داد و اصطلاحات مرتبط به آن را استخراج کرد به‌طوری که اگر دو مدرک یک اصطلاح مشابه داشته باشند، با توجه به نوع و نقش دستوری آنها می‌توان بین مفهوم آن اصطلاح در هر یک از دو مدرک تمایز قائل شد.

ب) روش بازخوردی بر قضاوت کاربران قبلی و مدل احتمالات استوار است بدان معنا که بر اساس توزیع آماریِ قضاوت کاربران قبلی دربارهٔ مفاهیم و اصطلاحات موجود در یک مدرک، قضاوت کاربران آینده را پیش‌بینی می‌کنند. پیش‌فرض روش بازخوردی این است که هر چه یک اصطلاح در یک مدرک قبلاً توسط کاربران بیشتری مرتبط تشخیص داده شده باشد، احتمال بیشتری وجود دارد که با منظور کاربر فعلی مرتبط باشد و لذا این اصطلاح احتمالاً در بردارندهٔ یکی از مفاهیم اصلی مدرک است. از این رو، اعتبار روش احتمالی به همخوانی بین تشخیص اصطلاحات یک مدرک توسط کاربران قبلی و فعلی بستگی دارد.

پ) در روش آماری، اساس کار توجه به فراوانی اصطلاحات و ارزش آنها برای یک مدرک است. فرض بر این است که بین تکرار یک اصطلاح در یک مدرک و منظور مدرک ارتباطی وجود دارد. این نکته را نباید از نظر دور داشت که از نگاه مدل استنتاجی، نقدی که بر روش‌های یاد شده وارد است این است که «منظور یک مدرک در فحوای آن نهفته نیست بلکه از یک مدرک در موقعیت‌های ارتباطی گوناگون، منظورهای گوناگونی برداشت می‌شود» (پارک[1]، ۱۹۹۶) که این مسأله کارآمدی نمایه‌سازی را زیر سؤال می‌برد. در نمایه‌سازی می‌توان کارآمدی[2] را شباهت بین قضاوت عاملان نمایه‌سازی (انسانی یا ماشینی) و کاربران اطلاعات در زمینهٔ مفاهیم و منظورهای یک مدرک دانست. از این رو، هیچ‌گاه نمی‌توان انتظار داشت که برداشت و نتایج نمایه‌سازی بین دو عامل انسانی و حتی یک عامل انسانی در دو زمان مختلف،

1. Park 2. effectiveness

همخوانی کامل وجود داشته باشد. به همین دلیل، در هر سامانهٔ نمایه‌سازی چالش در مورد توانمندی آن در تشخیص اصطلاحات مناسب پا بر جا است. برای کاهش این چالش، در ارزیابی توانمندی سامانهٔ پیشنهادی، بر سنجش همخوانی اصطلاحات حاصل از مدل با کلیدواژه‌های نویسندگان مقالات تکیه شده است.

هر زبان، ویژگی‌های منحصر به خود را دارد بنابراین در هر زبان مسائل منحصر به فردی برای پردازش متون وجود دارد. نمایه‌سازی خودکار نیز فعالیتی است که با پردازش متون سروکار دارد. الگوریتم‌های متفاوتی برای نمایه‌سازی خودکار وجود دارد که عموماً دارای سه مرحله‌اند: الف) استخراج واژه‌ها و عبارات ممنوعه؛ ب) ریشه‌یابی واژه‌ها برای یک‌دستی و حذف افزونگی؛ و پ) وزن‌دهی. در زبان فارسی، بر خلاف زبان انگلیسی و به علت دخیل بودن واژه‌های عربی، نگاه به ریشه‌یابی کاملاً متفاوت است. به همین دلیل، شاید بهتر باشد مرحلهٔ ریشه‌یابی را حذف و کارهای دیگری را جایگزین آن کرد. مسأله‌ای که منجر به طراحی و آزمون مدل دو مرحله‌ایِ حاضر شد، نامشخص بودن کارکرد عملی این مدل نظری در زبان فارسی بود. با توجه به مسائلی از این دست و بر اساس کارهای عملی، روشی برای نمایه‌سازی متون فارسی تجربه گردید که حاصل آن در ادامه می‌آید.

در نگاه به تاریخچهٔ نمایه‌سازی خودکار، دو مدل نظری مهم دیده می‌شود: مدل فضای برداری[1] و مدل احتمالی[2] (اندرسون و پرس-کاربالو[3]، ۲۰۰۱). در عمل، این دو مدل کاملاً از هم قابل تفکیک نیستند چون هر یک از این ایده‌ها از دیگری بهره گرفته است. در روش فضای برداری، هر مدرک یا اصطلاح، یک بردار هندسی در نظر گرفته می‌شود و بر اساس میزان اصطلاحات موضوعی مشترکِ دو مدرک یا تعداد مدارک مشترک بین دو اصطلاح موضوعی، شباهت بین دو بردار سنجیده می‌شود. در مدل احتمالی، احتمال میزان مرتبط بودن هر اصطلاح موضوعی با مدرک مورد نظر سنجیده می‌شود.

مسألهٔ مورد توجه در این نوشتار، روشن ساختن توانمندی مدل دو مرحله‌ای شکاف–گلچین[4] برای نمایه‌سازی متون فارسی است که می‌توان آن را در دستهٔ روش‌های احتمالی جای داد. این مدل حاصل تجربیات عملی پژوهشگر در زمینهٔ نمایه‌سازی با توجه به ویژگی‌های زبان

1. vector space model
2. probabilistic model
3. Anderson & Pérez-Carballo
4. break-cull

فارسی است و به‌گونه‌ای طراحی گردیده است که نیاز به استفاده از جدول‌های بانک اطلاعاتی برای محاسبات لازم را به حداقل می‌رساند. فنی که در پیاده‌سازی این مدل پیاده شده است، استفاده از زبان «عبارات الگودار[1]» در زبان‌های برنامه نویسی است که اکثر زبان‌های سطح بالا آن را پشتیبانی می‌کنند. این زبان بر الگوسازی استوار است. نمونه شبیه آن در جستجوی اطلاعات، بُرش دادن[2] عبارت جستجو است. مثلاً، ما اگر عبارت جستجو را به صورت «زندگ$» به یک سامانهٔ بانک اطلاعاتی بدهیم، عباراتی مانند «زندگی» و «زندگانی» جستجو خواهد شد. در این جا ما یک الگو ساخته‌ایم که ابتدای آن «زندگ» است و انتهای آن هر چیز می‌تواند باشد. زبان عبارات الگودار، امکان الگوسازی‌های متنوعی را به دست می‌دهد که کار روی متن را با سرعت اجرای بالا آسان می‌سازد.

اساس کار مدل حاضر این است که در مرحلهٔ شکاف، بر اساس برخی واژه‌ها و نشانه‌ها و با استفاده از زبان عبارات الگودار، شکاف‌هایی در متن ایجاد می‌شود. احتمال دارد واژه‌ها و عباراتی که بین این شکاف‌ها قرار می‌گیرند، اصطلاحاتی باشند که دربردارندهٔ مفهوم هستند. با فن شکاف‌دادن می‌توان اصطلاحاتی را استخراج کرد که در قالب پیش‌فرض‌های مطرح در مقالهٔ کاگرا[3] (۱۹۹۶) می‌گنجند. این پیش‌فرض‌ها عبارتند از:

۱. هر واژه‌ای که در یک مدرک وجود دارد، احتمالاً یک اصطلاح موضوعی برای آن مدرک است؛
۲. هر واژه‌ای که در یک مدرک دارای فراوانی زیاد است، احتمالاً یک اصطلاح موضوعی برای آن مدرک است؛
۳. دو یا چند واژه‌ای که متداولاً با هم می‌آیند، احتمالاً یک اصطلاح عبارتی هستند؛
۴. وقتی که دو اصطلاح غالباً در کنار هم بیایند، تشکیل یک اصطلاح عبارتی مرکب می‌دهند؛
۵. وقتی که یک واژه غالباً با اصطلاحات مهم می‌آید، خودِ آن واژه، یک اصطلاح ساده است.
۶. وقتی که یک واژه غالباً با یک اصطلاح مهم می‌آید، با هم تشکیل یک اصطلاح عبارتی مرکب می‌دهند.

1. Regular expression 2. truncation 3. Kageura

پس از مرحله شکاف، سعی می‌شود با ارائهٔ روش‌ها و توابع ریاضی از میان اصطلاحات استخراج شده، جامع‌ترین و دقیق‌ترین آنها گلچین شوند که در این مرحله نیز از زبان عبارات الگودار بهره گرفته می‌شود.

معرفی مدل

همان‌گونه که مختصراً ذکر شد، در مدل پیش رو برای تشخیص اصطلاحات کلیدی در متون فارسی دو مرحلهٔ اصلی در نظر گرفته می‌شود و تلاش بر این است که نتایج حاصل با کلیدواژه‌های نویسندگان نزدیک باشد:

۱. تشخیص و خارج ساختنِ واژه‌ها یا عباراتی که به احتمال زیاد خود حاوی مفهوم نیستند و در نتیجه، تعیین واژه‌ها و عباراتی که احتمالاً حاوی مفهوم هستند (مرحلهٔ ایجاد شکاف)؛

۲. تشخیص و تعیین واژه‌ها یا عباراتی که به احتمال زیاد اصطلاحات جامع و دقیق هستند (مرحلهٔ گلچین کردن مرتبط‌ترین اصطلاحات).

هدف مرحلهٔ اول، تعیین و علامتگذاری ابتدا و انتهای اصطلاحات معنی‌دار در متن است. پیش‌فرض این است که واژه‌ها یا عباراتی که بین دو واژه یا عبارت بدون مفهوم قرار بگیرند، احتمالاً اصطلاحاتی مفهوم‌دار هستند. جملهٔ زیر، برگرفته از کتاب «آشنایی با علم‌سنجی» نوشتهٔ عبدالرضا نوروزی چاکلی می‌تواند این گفتار را به تصویر بکشد:

(۱) اصل کمترین کوشش به این معناست که یک شخص می‌کوشد مشکلات فوری و به احتمال، دشواری‌های آینده‌اش را از طریقی که به کمترین تلاش نیاز داشته باشد حل کند.

جملهٔ بالا با حذف کلمات و عبارات بدون مفهوم، به شکل ذیل درمی‌آید:

(۲) اصل کمترین کوشش * * *** شخص * مشکلات فوری * * ** دشواری‌های آینده* * * * * * کمترین تلاش ******

مورد (۲) انعکاسی از مورد (۱) است با این تفاوت که واژه‌ها و عبارات بدون مفهوم از آن حذف گردیده و به جای آنها علامت «*» قرار گرفته است. اگر این دو نمونه را با هم مقایسه کنیم، می‌بینیم که حروف اضافه، ضمایر، علائم و اَفعال از مورد (۱) حذف گردیده و به جای آن علامت ستاره قرار گرفته است. احتمال می‌رود آن اصطلاحاتی که در ابتدا یا انتهای متن هستند یا اصطلاحاتی که بین دو ستاره قرار می‌گیرند، حاوی مفهوم باشند و بنابراین می‌توان از

میان آنها مرتبط‌ترین اصطلاحات را با فنونی که در ادامه می‌آید، گلچین کرد. با این توضیح، در ادامه بر بخش‌هایی از متن در متون فارسی توجه می‌کنیم که احتمالاً حاوی مفهوم نیستند و می‌توانند نقش شکاف‌دهنده داشته باشند.

شکاف‌دهنده‌های متن

برخی از واژه‌ها یا عبارات موجود در متون، نمی‌توانند کلیدواژه یا اصطلاح قرار بگیرند که عبارتند از:

۱. علائم: از علائم نقطه‌گذاری گوناگونی در متن استفاده می‌شود. این علائم اولین جاهایی هستند که می‌توان از آنجا متن را شکاف داد. علائمی چون نقطه، پرانتز، ویرگول، نقطه‌ویرگول و بسیاری از علائم دیگر.

۲. اَفعال. در نمایه‌سازی خودکار معمولاً توجه اصلی به عبارات و کلمات اسمی است و گمان می‌رود که این واژه‌ها به احتمال زیاد حاوی مفهوم هستند، لذا اَفعال را می‌توان به‌عنوان شکاف‌های متن تلقی کرد. تعیین و تشخیص اَفعال فارسی را می‌توان به صورت‌های متفاوتی انجام داد. از نمونه‌های آن، روش برنجیان (۱۳۹۰) برای ریشه‌یابی ماضی و مضارع اَفعال ناگذر در زبان فارسی است. کار مولودی (۱۳۹۰) دربارهٔ فعل مرکب و معیارهای صوری برای تشخیص آن نمونه‌ای دیگر است. در زمینهٔ تشخیص اَفعال فارسی کارهای نرم‌افزاری نیز انجام شده است که می‌توانند مرجع عملی خوبی باشند. یکی از قدیمی‌ترین آنها، توسط دانش‌کار آراسته (۱۳۸۳) صورت گرفته است که به طراحی نرم‌افزار تشخیص فعل در زبان فارسی پرداخته است. در اثر حاضر، روش‌های گوناگونی آزموده شد که از جملهٔ آنها، صرف فعل با تعیین بُن زمان آن بود. ویژگی این روش این است که اگر بُن یک فعل برای یک زمان مشخص (مثلاً گذشتهٔ استمراری) داده شود، تمام صیغه‌های آن صرف خواهد شد. این باعث می‌شود که برای هر فعل لازم نباشد که تمام صیغه‌های آن صرف شود و فرایند برنامه‌نویسی و اضافه کردن اَفعال جدید به فهرست اَفعال، آسان باشد. این الگوریتم نسبت به سایر الگوریتم‌ها سریع‌تر است اما یک روش سریع‌تر دیگر، صرف یک‌به‌یک تمام صیغه‌ها و زمان‌های هر فعل در

پیکرهٔ کدهای برنامه است. گر چه این روش در مرحلهٔ اجرای برنامه، از لحاظ زمانی بسیار سریع‌تر است اما در هنگام نوشتن برنامه وقت زیادی گرفته می‌شود.

در همهٔ روش‌هایی که آزمون گردید، اَفعال فارسی به دو دسته تقسیم شدند: اَفعال کمکی و امثال آن، و اَفعال اصلی. تشخیص اَفعال کمکی ساده است، زیرا از نظر تعداد محدود هستند. در این گونه اَفعال، فعل به محض پیدا شدن در متن به محل شکاف تبدیل می‌شود. مثلا در جملهٔ «من می‌خواهم امروز به دانشگاه بروم»، فعل کمکی «می‌خواهم» به یک شکاف تبدیل می‌شود و همانند مورد ۲ به ستاره یا هر علامت دیگری به عنوان نشانگر شکاف بدل می‌گردد.

در کنار سایر تقسیم‌بندی‌هایی که برای اَفعال اصلی در فارسی وجود دارد، می‌توان آنها را به دو دستهٔ پرکاربرد و کم‌کاربرد تقسیم کرد. آنچه که باید بر آنها تمرکز شود، اَفعال پر کاربرد است و دست‌کم معرفی و صرف این اَفعال در برنامه کافی است زیرا تجربه نشان داده است که اگر هر یک از اَفعال کم‌کاربرد، به اشتباه به‌عنوان اصطلاح بین دو شکاف قرار بگیرند، در ادامه فرایند نمایه‌سازی خودبه‌خود از گردونهٔ نهایی اصطلاحات گلچین شده حذف می‌گردند. این بدان معنا نیست که نباید اصلاً به اَفعال کم‌کاربرد توجه داشت بلکه منظور این است که حداقل به اَفعال پرکاربرد توجه شود. می‌توان به اَفعال پرکاربرد زبان فارسی که در پژوهش ساده، بِکلی و تقوا[1] (۲۰۰۳) استخراج شده است، تکیه کرد که عبارتند از: کردن، شدن، گردیدن، آوردن، بردن، خوردن، گرفتن، نشستن، یافتن و ... . معمولاً واژه‌ای که قبل از این اَفعال قرار می‌گیرد نیز می‌تواند حذف شود. مثلاً در جملهٔ «من امروز زمین خوردم»، «زمین» که قبل از «خوردم» آمده است می‌تواند حذف شود. لذا حداقل کاری که باید صورت گیرد آن است که وقتی که یک فعل اصلی پیدا شد، واژهٔ قبل از آن هم به محل شکاف تبدیل شود. اما هر چه دقت در شناخت اَفعال اصلی یک کلمه‌ای و چندکلمه‌ای بیشتر باشد، نتیجه‌ای مطلوب‌تر حاصل می‌شود. نکته‌هایی که اشاره شد، حداقل کاری است که باید در مورد اَفعال صورت گیرد تا در نهایت اصطلاحاتی نسبتاً مطلوب گلچین شود.

1. Sadeh, Beckley & Taghva

۳. واژه‌ها و عبارت‌های ممنوعه: بخش دیگری از یک متن که می‌تواند به‌عنوان شکاف‌دهنده در نظر گرفته شود، واژه‌ها و عبارت‌های ممنوعه هستند. از این جهت به آنها ممنوعه گفته می‌شود که در جستجوهای معمول برای بازیابی اطلاعات برای آنها ارزش مفهومی قائل نمی‌شوند و نظر بر این است که در بردارندهٔ مفهوم نیستند و لذا باید از به‌کار بردن آنها در عبارات جستجو پرهیز کرد. احتمال حضور این واژه‌ها و عبارت‌ها در یک مدرک مرتبط با نیاز اطلاعاتی یک فرد به اندازهٔ یک مدرک غیرمرتبط است. واژه‌ها و عبارت‌های ممنوعه را به صورت‌های متفاوت دسته‌بندی کرده‌اند اما در حین پیاده‌سازی الگوریتم، مشاهده شد که در زبان فارسی می‌توان آنها را به چهار دستهٔ اصلی تقسیم کرد:

الف. مواردی که در هر حال و همیشه ممنوعه هستند، مانند بسیاری از حروف اضافه و ضمایر. بر اساس تجربه و نتایج حاصل، معلوم گردید که یک فهرست قابل قبول از این نوع، باید دست‌کم حاوی هشتصد واژه و عبارت ممنوعه باشد.

ب. مواردی که بهتر است از اول اصطلاح موضوعی حذف شود. این اصطلاحات، مواردی هستند که در جستجوی اطلاعات نیز معمولاً از ابتدای عبارت جستجو حذف می‌شوند، مانند: تأثیر، نقش، تعیین و تبیین.

پ. مواردی که نمی‌تواند در انتهای اصطلاح قرار گیرد، مانند برخی حروف اضافه که جزو مورد الف قرار نمی‌گیرد. «و»، «از»، «با» و ... نمونه‌هایی از این حروف‌اند که می‌توانند جزئی از یک اصطلاح باشند.

ت. مواردی که گاهی جزو اصطلاح هستند و گاهی می‌توانند شکاف‌دهنده باشند. واژه‌های «و»، «از»، «به»، و «با» از اصلی‌ترین آنها هستند. مثلاً در اصطلاح «آموزش و پرورش»، واژهٔ «و» جزئی از اصطلاح است ولی در عبارت «دزدان دریایی و امنیت دریا»، «و» فقط می‌تواند یک شکاف‌دهنده باشد چون «دزدان دریایی»و «امنیت دریا» دو اصطلاح مجزا هستند. البته مورد (ت) در مرحلهٔ تعیین واژه‌ها و عبارات ممنوعه مورد توجه قرار نمی‌گیرد، بلکه در مرحلهٔ همگن‌سازی حالت‌های مختلف یک عبارت اهمیت پیدا می‌کند که در بخش‌های بعدی به آن اشاره خواهد شد.

نکته قابل توجه دربارهٔ این نوع واژه‌ها و عبارات، رسم‌الخط گوناگون آنها است که به صورت‌های گوناگون نوشته می‌شوند، مثلاً عبارت ممنوعهٔ «به طوری که» ممکن است به صورت «بطوریکه»، «به طوریکه»، «به‌طوری‌که» نوشته شود. بنابراین، توجه به حالت‌های نوشتن یک عبارت ممنوعه ضروری است.

تشخیص و تعیین اصطلاحات جامع و دقیق (گلچین کردن)

همانند مورد (۲)، واژه‌ها و عباراتی را که پس از مرحلهٔ شکاف باقی می‌مانند، «اصطلاحات بین شکافی» می‌نامیم. احتمال این که این اصطلاحات حاوی مفهوم باشند، بسیار بالا است ولی مسلماً تعداد آنها زیاد است و باید مرتبط‌ترین‌ها را گلچین کرد. در اینجا دو سؤال اساسی پیش می‌آید: کدام‌یک از اصطلاحات بین شکافی مرتبط‌ترند؟ چه تعداد از آنها باید گلچین شود تا جامعیت و مانعیت نمایه‌سازی حفظ شود؟

وزن‌دهی

یکی از کارهای مقدماتی برای تعیین مرتبط‌ترین‌ها، وزن دادن به اصطلاحات است. روش‌های متعددی برای وزن‌دهی وجود دارد که یکی از مشهورترین آنها روش تی.اف-آی.دی.اف[1] یا «فروانی اصطلاح-عکس فراوانی مدرک» است. برای پیاده‌سازی این روش به مجموعه‌ای از مدارک نیاز است. ابتدا هر واژه یا عبارتی که در متن در بردارندهٔ مفهوم باشد به‌عنوان یک اصطلاح احتمالی در نظر گرفته می‌شود و فراوانی آن در متن مورد نمایه‌سازی محاسبه می‌شود. هر چه فراوانی یک اصطلاح در متن بیشتر باشد، ارزش آن اصطلاح در آن متن بیشتر است که به آن «وزن محلی» می‌گویند. علاوه بر آن، ارزش آن اصطلاح در همهٔ مدارک سنجیده می‌شود زیرا هر چه یک اصطلاح در مدارک بیشتری دیده شود، ارزش آن پایین‌تر است. به این فراوانی، «وزن سراسری» اصطلاح گفته می‌شود. وزن نهایی یک اصطلاح، حاصل ضرب دو وزن محلی و سراسری است. توابع مختلفی برای تعیین وزن اصطلاح بر اساس این روش وجود دارد که عموماً با تفاوت‌های جزئی بر نرمال کردن نتیجه تأکید دارند. یکی از این فرمول‌ها که توسط سالتن و باکلی[2] (۱۹۸۸) ارئه شده، به شکل زیر است:

1. Term Frequency – Inverse Document Frequency (TF-IDF)
2. Salton & Buckley

$$wd = fw, d * \log (|D|/fw, D) \quad (2)$$ (الف

در فرمول بالا، wd برابر با وزن اصطلاح و fw, d فراوانی یا تعداد دفعاتی است که اصطلاح مورد نظر در مدرک d ظاهر شده است و|D|تعداد مدارکی است که در مجموعه وجود دارد و fw, D نیز تعداد مدارکی است که اصطلاح w را در خود دارند.

روش وزن‌دهی بالا، همیشه به مجموعه‌ای از مدارک وابسته است و در صورت افزوده شدن یک مدرک دیگر به آن مجموعه باید تمامی مدارک مجدداً نمایه‌سازی شود. از این رو، این روش بیشتر مناسب مجموعه‌های ثابت است. در مدل شکاف- گلچین می‌توان اصطلاحات بین شکافی هر مدرک را بدون نیاز به مجموعه‌ای از مدارک دیگر وزن‌دهی کرد.

مسائل پیش روی وزن‌دهی (یک‌دست‌سازی قبل از وزن‌دهی)

چون وزن‌دهی تابع فراوانی است، قبل از آن باید همگن‌سازی یا یک دست‌سازی اصطلاحات صورت گیرد تا فراوانی واقعی محاسبه گردد. مثلاً در عمل دو عبارت «سامانه اطلاعات مدیریت» و «سامانه‌های اطلاعات مدیریت» یک اصطلاح واحد هستند ولی یکی جمع و دیگری مفرد است و لذا باید یکی را برگزید. در زیر چهار مورد که باعث به وجود آمدن چنین حالاتی در فارسی می‌شود ذکر گردیده و برای آن راه‌حلی پیشنهاد شده است:

۱. واژه‌های احتمالاً ممنوعه[1]: قبلاً دیدیم که برخی واژه‌ها یا عبارت‌ها در جاهایی واژهٔ ممنوعه و در جایی دیگر بخشی از یک اصطلاح هستند. معروف‌ترین آنها «و»، «از»، «به» و «با» بودند. برای درک بهتر این مطلب، دو عبارت «افسردگی دانشجویان» و «افسردگی در دانشجویان» را در نظر بگیریم. این دو عبارت عملاً یکی هستند ولی یکی با حرف اضافه و دیگری بدون حرف اضافه. در این موارد، اولویت را به عبارات بدون حرف اضافه می‌دهیم یعنی فراوانی هر دو حالت را در نظر می‌گیریم و اگر تعداد تکرار عبارت بدون حرف اضافه با تکرار عبارت با حرف اضافه در متن مورد نظر برابر یا بزرگ‌تر بود، عبارت بدون حرف اضافه را به عنوان اصطلاح می‌پذیریم و در صورتی که تکرار عبارت با حرف اضافه در متن بیشتر بود، احتمالاً آن اصطلاح با حرف اضافه معمول‌تر است و لذا اصطلاح را با حرف اضافه می‌پذیریم.

1. stopwords

۲. عبارات جایگشت شده[1]: برخی عبارات هستند که حاوی واژه‌های احتمالاً ممنوعه هستند ولی در متن گاهی تقدم و تأخر بخش‌های قبل و بعد آنها جابه‌جا می‌شود که ما به این حالت اصطلاحاً «جایگشت عبارت» می‌گوییم. مثلاً دو عبارت «حذر و تعویذ» و «تعویذ و حذر» عملاً یکی هستند که در اولی «حذر» و در دومی «تعویذ» مقدم شده است. لذا عباراتی که حاوی واژه‌های احتمالاً ممنوعه هستند، باید آنها را جایگشت داد و در صورتی که هر دو حالت وجود داشته باشد، موردی که فراوانی بیشتری در متن دارد برگزینیم و حالت دیگر را در کل متن به حالت عبارت برگزیده درآوریم.

۳. حالت جمع و مفرد: تشخیص جمع فارسی واژه‌ها در متن، نسبتاً ساده است زیرا جمع فارسی معمولاً با پسوند «ها» و «ان»ساخته می‌شود. بنابراین هر واژه‌ای که با این رشته‌ها ختم شود، ابتدا «ها» یا «ان» را حذف می‌کنیم تا ببینیم آیا در متن عباراتی مشابه ولی بدون ختم شدن به این دو پسوند وجود دارد یا خیر. سپس فراوانی آن عبارت را با پسوند و بدون پسوند محاسبه می‌کنیم. اگر تعداد بدون پسوند در متن بیشتر بود، حالت مفرد را می‌پذیریم و تمام جمع‌ها را به مفرد تبدیل می‌کنیم و اگر حالت جمع، برابر یا بیشتر از حالت مفرد بود، حالت جمع را بر می‌گزینیم و تمام عبارت‌های مفرد را در متن به جمع تبدیل می‌کنیم. مشکل اساسی، تشخیص جمع واژه‌های زبان عربی موجود در فارسی است. جمع‌های سالم عربی مانند آنچه که بیان شد قابل تشخیص است، اما تشخیص واژه‌هایی با جمع مکسر، کمی پیچیده‌تر است اما با تمرکز بر ریشهٔ کلمات (ف ع ل) و حروف مزید آنها، در کنار شناخت صیغه‌های صَرف اسامی و واژه‌های عربی(مثل اسم فاعل، مفعول) این کار میسر می‌شود.

۴. یاء مضاف یا موصوف و قید. واژه‌هایی هستند که وقتی در حالت مضاف یا موصوف قرار می‌گیرند در انتهای آنها حرف «ی» می‌آید؛ مانند «آرزوی محال» یا «گفتاری زیبا». برای حل این مسأله، می‌توان فراوانی اصطلاحات بین شکافی را که واژه‌های آنها با «ی» ختم می‌شود، یک بار بدون «ی» شمرد. اگر نتیجهٔ حاصل نشان داد که فراوانی بدون «ی» بیشتر است، حالت بدون ختم به «ی» را می‌پذیریم و «ی» را از عبارت دیگر حذف می‌کنیم.

1. permuted

محاسبهٔ وزن

در روش ابداعی، چند پارامتر برای وزن‌دهی مورد توجه قرار گرفت و پس از آزمون‌های فراوان، ضرایبی برای آنها تعیین گردید که عبارتند از: الف) فراوانی قرار گرفتن اصطلاح بین دو شکاف؛ ب) فراوانی قرار گرفتن هر اصطلاح به‌عنوان جزئی از یک اصطلاح بزرگ‌تر؛ پ) میانگین فراوانی هر یک از واژه‌های تشکیل‌دهندهٔ اصطلاح مورد نظر؛ و نهایتاً ت) تعداد واژه‌های تشکیل‌دهندهٔ آن اصطلاح.

اگر فرض کنیم که همانند مورد (۲) اصطلاحات بین شکافی همان‌هایی هستند که بین دو ستاره قرار گرفته‌اند یا در ابتدا و انتهای متن آمده‌اند، آن‌گاه منظور از فراوانیِ قرار گرفتن یک اصطلاح بین دو شکاف، تعداد تکرار یک اصطلاح در حالتی است که دو طرفش ستاره باشد. فراوانی بین شکافی، یکی از مهمترین شاخص‌های محاسبهٔ وزن یک اصطلاح است. گاه یک اصطلاح می‌تواند هم به صورت کاملاً بین‌شکافی و هم به‌عنوان بخشی از یک اصطلاح بین‌شکافی دیگر باشد. مثلاً ممکن است در یک متن، دو اصطلاح بین شکافی *مدیریت اطلاعات* و *سامانه‌های مدیریت اطلاعات* وجود داشته باشد که «مدیریت اطلاعات» علاوه بر این که در جاهایی از متن مستقلاً یک اصطلاح بین‌شکافی است، در جایی دیگر جزئی از یک اصطلاح بین‌شکافی دیگر به نام «سامانه‌های مدیریت اطلاعات» باشد. بنابراین، «مدیریت اطلاعات» علاوه بر این که یک فروانی بین‌شکافی دارد، یک فراوانی دیگر دارد که اصطلاحاً آن را «فراوانی جزئی» می‌نامیم. هر اصطلاح می‌تواند از یک یا چند واژه تشکیل شده باشد، مثلاً «سامانه‌های مدیریت اطلاعات» از سه واژهٔ «سامانه‌ها»، «مدیریت» و «اطلاعات» تشکیل شده است که هر کدام از این واژه‌ها به تنهایی دارای فراوانی مشخصی هستند. میانگین فراوانی واژه‌های تشکیل‌دهندهٔ یک اصطلاح می‌تواند پارامتری دیگر باشد که اصطلاحاً به آن «فراوانی واژه‌ای» می‌گوییم که بیانگر تعداد واژه‌های تشکیل‌دهندهٔ یک اصطلاح است. این پارامتر از آن جهت اهمیت دارد که هر چه تعداد واژه‌های تشکیل‌دهندهٔ یک اصطلاح بیشتر باشد، فراوانی بین‌شکافی آن کمتر خواهد بود. برای جبران این مسأله باید به اصطلاحاتی که تعداد واژه‌های تشکیل‌دهندهٔ آن بیشتر است، ارزش بیشتری بدهیم. به این مقدار، اصطلاحاً «ارزش تعداد واژه‌ها» می‌گوییم. پس از آزمون‌های مختلف، مشخص گردید که هر چه تعداد واژه‌های

تشکیل‌دهندهٔ یک اصطلاح افزایش یابد، نقش و اهمیت تعداد واژه‌های تشکیل‌دهنده کمتر می‌شود و لذا تصمیم گرفته شد تا از یک تابع لگاریتمی استفاده شود:

$$V = Ln(X) + 1$$

در فرمول بالا، V برابر است با ارزش تعداد واژه‌ها، Ln نشانهٔ لگاریتم طبیعی، و X تعداد واژه‌های تشکیل‌دهندهٔ یک اصطلاح است. حال بر اساس فرمول زیر می‌توان این چهار پارامتر را ترکیب کرد و برآیند آن را به عنوان وزن اصطلاح نامید. پس از آزمایش‌های فراوان روی مدارک مختلف، معلوم گردید که فراوانی بین‌شکافی ارزشی ۲ برابر سایر فراوانی‌ها دارد، به همین جهت، این فراوانی با ضریب ۲ در نظر گرفته شد:

$$W = (2B + P + M) \times V$$

در فرمول بالا، W برابر است با وزن اصطلاح، B برابر است با فراوانی بین‌شکافی، P برابر است با فراوانی جزئی و M معادل فراوانی واژه‌ای و در نهایت V نشانگر ارزش تعداد واژه‌های تشکیل‌دهندهٔ یک اصطلاح است. بر اساس این فرمول، وزن هر اصطلاح عددی بالاتر از صفر خواهد بود. جالب است که اگر اصطلاحات را بر اساس وزن آنها از زیاد به کم تنظیم کنیم یک رابطهٔ لگاریتمی ایجاد خواهد شد. البته این مطلب نیاز به پژوهش مستقل دارد.

تعداد اصطلاحات گلچین شده (تعیین آستانه)

قبلاً این پرسش را مطرح کردیم که چه تعداد از اصطلاحات بین‌شکافی باید به عنوان کلیدواژه‌ها یا اصطلاحات نهایی گلچین شوند تا جامعیت و مانعیت نمایه‌سازی حفظ شود. اصطلاحات استخراج شده به روش‌های مختلف وزن‌دهی می‌شوند، اما دشواری اینجا است که کدام‌ها و چه تعداد را برگزینیم. به همین دلیل، مباحثی مثل محدوده یا آستانهٔ برش، یعنی بالاترین و پایین‌ترین حد انتخاب اصطلاحات از دیر باز در آثار کسانی مثل لوهن[1] (۱۹۵۸) مطرح است. از آنجا که در نمایه‌سازی باید بین جامعیت و مانعیت یک تعادل معقول صورت گیرد، در برخی روش‌ها مثل تی.اف-آی.دی.اف، اگر مبنا را اصطلاحاتی بگیریم که بیشترین وزن را دارند تعادل بین جامعیت و مانعیت به هم می‌خورد. بنابراین، آستانهٔ بالا و پایین در نظر گرفته می‌شود که مثلاً اصطلاحات بین محدودهٔ وزن ۰/۲ و ۰/۵ برگزیده شوند که تعیین نقطه آغازین و پایانی این محدوده نیز خود چالش برانگیز است.

1. Luhn

در روش حاضر، علاوه بر تعیین تعداد اصطلاحات گلچین‌شدهٔ نهایی، به یک فهرست گلچین شده اولیه نیاز است. این فهرست از میان کل اصطلاحات بین شکافی و در فرایند تشخیص اصطلاحات افزونه ساخته می‌شود. بنابراین، در ادامه ابتدا دربارهٔ ساختن فهرست گلچین‌شدهٔ اولیه صحبت خواهد شد و سپس به نحوهٔ تعیین تعداد اصطلاحات گلچین‌شده خواهیم پرداخت. چون ساخت فهرست گلچین‌شدهٔ اولیه با مسألهٔ رفع افزونگی مرتبط است، این بحث تحت عنوان افزونگی مطرح می‌شود.

افزونگی و تعیین تعداد اصطلاحات گلچین‌شدهٔ اولیه

یکی از مسائل نمایه‌سازی در هر دو حالت دستی و خودکار، مسألهٔ افزونگی است که در مقدمه به آن اشاره شد. در روش پیشنهادی، الگوریتم کنار گذاشتن اصطلاحات افزونه بر سه پیش‌فرض استوار گشت. اگر فهرست اصطلاحات بین‌شکافی، بر اساس وزن آنها تنظیم شود، آنگاه پیش‌فرض‌ها عبارت خواهند بود از:

۱. اصطلاحی که وزن بیشتری دارد، بر اصطلاحات بعدی خود مقدم است و در فهرست اولیهٔ گلچین شده‌ها قرار می‌گیرد.
۲. اصطلاحاتی که بعد از یک اصطلاح قرار می‌گیرند و به‌عنوان جزئی از آن اصطلاح محسوب می‌شوند، جزو فهرست اولیهٔ گلچین شده‌ها قرار نمی‌گیرد.
۳. اصطلاحاتی که بعد از یک اصطلاح قرار می‌گیرند و آن اصطلاح را به‌عنوان جزئی از خود در بر دارند، جزو فهرست اولیهٔ گلچین شده‌ها قرار نمی‌گیرد.

برای روشن‌تر شدن مطلب، مثال می‌زنیم. فرض کنیم که با فنون یاد شده، یک مدرک را نمایه‌سازی کرده و فهرستی از اصطلاحات بین‌شکافی به دست آورده‌ایم که این اصطلاحات به ترتیب وزنشان به شکل زیر تنظیم شده است:

سامانه‌های مدیریت اطلاعات، مدیریت اطلاعات، اطلاعات بیمارستانی، ساماندهی اطلاعات، ساماندهی اطلاعات بیمارستانی، پرستاران، مدیریت

طبق پیش‌فرض ۱، اصطلاح «سامانه‌های مدیریت اطلاعات» چون وزن بیشتری داشته است، بر سایر اصطلاحات این فهرست مقدم است. لذا آن را برمی‌گزینیم و در فهرست اولیهٔ گلچین‌شده‌ها قرار می‌دهیم. طبق پیش‌فرض ۲، «مدیریت اطلاعات» چون جزئی از اصطلاح

گلچین‌شدهٔ «سامانه‌های مدیریت اطلاعات» است، در فهرست گلچین‌شده‌ها قرار نمی‌گیرد. همین‌طور اصطلاح «مدیریت» طبق همین پیش‌فرض، از فهرست اصطلاحات گلچین شده بیرون می‌ماند. چون تکلیف اصطلاحات قبل از اصطلاح «اطلاعات بیمارستانی» مشخص شده است، در حال حاضر در میان اصطلاحات باقیمانده، بیشترین وزن را دارد. لذا طبق پیش‌فرض ۱ به فهرست گلچین‌شده‌ها منتقل می‌شود. در میان اصطلاحات باقیمانده، «ساماندهی اطلاعات بیمارستانی»، طبق پیش‌فرض ۳ نمی‌تواند جزو فهرست گلچین‌شده‌ها قرار بگیرد زیرا بعد از اصطلاح گلچین‌شدهٔ «ساماندهی اطلاعات» قرار گرفته است و جزئی از آن را در بر دارد. در آخر، اصطلاح «پرستاران» باقی می‌ماند که بر اساس پیش‌فرض ۱ به فهرست گلچین‌شده‌ها منتقل می‌گردد. در نهایت فهرست اولیهٔ گلچین‌شده‌ها به صورت زیر درمی‌آید:

سامانه‌های مدیریت اطلاعات، اطلاعات بیمارستانی، ساماندهی اطلاعات، پرستاران

نکته قابل توجه این است که در این روش به‌طور خودکار تصمیم گرفته می‌شود که از میان حالت عام و خاص یک اصطلاح، کدام را باید ترجیح داد. به فرض، در مثال بالا از میان سه اصطلاح «سامانه‌های مدیریت اطلاعات»، «مدیریت اطلاعات» و «مدیریت»، طبق الگوریتم، اصطلاح اولی که خاص‌تر است ترجیح داده شده است اما در مقابل، از میان «ساماندهی اطلاعات» و «ساماندهی اطلاعات بیمارستانی» اصطلاح اولی که عام‌تر است، ترجیح داده شده است.

تابع تعیین تعداد اصطلاحات گلچین‌شدهٔ نهایی

در روش حاضر، آستانهٔ بالایی برگزیدن اصطلاحات همیشه از اصطلاحی شروع می‌شود که بیشترین وزن را دارد. به عبارتی، اگر ما اصطلاحات گلچین‌شدهٔ اولیه را بر اساس وزنشان از زیاد به کم تنظیم کنیم، اصطلاحی که در ابتدا قرار می‌گیرد، نقطهٔ آغازین گلچین کردن نهایی است. از آنجا که تعداد کلیدواژه‌ها باید در حد معقولی باشد، این گلچین کردن در کجا باید به پایان برسد؟ توابع گوناگونی مورد آزمایش قرار گرفت تا بتوان نقطهٔ پایانی گزینش اصطلاحات موضوعی را تعیین کرد. مبنای صحت این توابع، محاسبهٔ همخوانی اصطلاحات حاصل از این توابع با کلیدواژه‌های نویسندگان بود که این کلیدواژه‌ها پس از چکیدهٔ مقالات می‌آید. فرمول تابع نهایی مورد قبول مانند زیر است:

$$n = \frac{b}{d/5}$$

در فرمول بالا (فرمول ۳)، n برابر است با تعداد اصطلاحات بین‌شکافی که باید از ابتدای فهرست گلچین شوند، b فراوانی اصطلاحات بین‌شکافی و d تعداد واژگان اصطلاحات بین‌شکافی است. منظور از تعداد واژگان اصطلاحات بین‌شکافی، فراوانی متمایز[1] این اصطلاحات است. تابع بالا با این فرض تعیین گردیده است که پس از عملیات نمایه‌سازی خودکار، امکان کنترل نمایه‌سازی از طریق واسطهٔ انسانی ممکن است. به همین خاطر عدد ثابت آن برابر با ۵ در نظر گرفته شده است تا حداکثر ممکن گلچین شود. به همین دلیل، گاهی یک یا دو اصطلاح گلچین‌شدهٔ انتهایی، دلچسب نیستند. برای این‌که احتمال استخراج کلیدواژه‌های دقیق بیشتر شود، می‌توان مقدار عدد ثابت در فرمول را افزایش داد اما با این کار احتمال از دست دادن اصطلاحات لازم نیز وجود دارد که نبودن آنها ممکن است جامعیت نمایه‌سازی را کاهش دهد.

مسائل قبل از پیاده‌سازی الگوریتم

روش حاضر در صورتی قابل پیاده‌سازی است که بتوان یک متن الکترونیکی را به رشته تبدیل کرد. منظور از رشته، مجموعه‌ای از کاراکترهای حرفی، عددی یا علائم یا ترکیبی از آنها است. مثلاً، عبارت «علی ساعت ۲ (به خانه) می‌رود»، یک رشته است که از چند حرف گوناگون، عدد ۲، فاصله، و پرانتز تشکیل شده است. برای پیاده‌سازی روش حاضر، علاوه بر مسألهٔ تبدیل متون الکترونیکی به رشته، مسألهٔ رسم‌الخط فارسی نیز باید مورد توجه قرار گیرد:

۱. مسائل تبدیل فایل به رشته

مدارک و نوشته‌های الکترونیکی به فرمت‌های گوناگونی عرضه می‌شوند. معروف‌ترین فرمت‌های موجود عبارتند از txt، rtf، doc، docx، pdf و html. مناسب‌ترین فرمت برای این روش، txt است. فراخوانی این نوع فایل‌ها و ریختن مطالب آنها روی یک رشته، نسبت به سایر فرمت‌ها ساده‌تر است. فایل‌های rtf، doc، docx و html یک متن را به صورت فرمت‌شده درون خود جای داده‌اند. از این رو، علاوه بر متن، حاوی محتویات دیگری نیز هستند که برای تبدیل

1. distinct frequency

و ریختن آنها درون یک رشته باید متن را از محتویات دیگر جدا ساخت. ضمناً، در سه نوع اول، متن به صورت کدهایی خاص ذخیره گردیده است که کدگشایی آنها، تنها با نرم‌افزارهای مربوط ممکن است. مثلاً برای خواندن یک فایل با فرمت docx باید حتماً نرم‌افزار Microsoft Word 2003 و بالاتر وجود داشته باشد تا بتوان فایل حاوی نوشته‌ها را کدگشایی و قابل نمایش کرد. در زبان‌های برنامه‌نویسی امروزی معمولاً امکانات و توابعی برای تبدیل این نوع فایل‌ها به متن ساده وجود دارد و همین‌طور، این امکان وجود دارد که برنامه‌های موجود را از جاهای دیگر گرفت و به امکانات آنها افزود. این امکانات و توابع را می‌توان به دو دسته تقسیم کرد:

الف. آنهایی که در صورت نصب‌بودن نرم‌افزار مربوطه، با کمک آنها می‌توان متن را استخراج و به رشته تبدیل کرد. مثلاً، در صورت نصب Microsoft Office Word، امکانات زبان برنامه‌نویسی اجازه می‌دهد که متن درون یک فایل را با فرمت doc استخراج کرد. در این حالت محاسن و معایبی وجود دارد، حُسن آن عدم نیاز به پیوست فایل‌های جانبی به فایل اجرایی اصلی برای اجرای برنامه است و عیب آن، وابسته بودن اجرای برنامه به نصب نرم‌افزارهای دیگر است. البته چون روی اکثر رایانه‌های امروزی این برنامه‌ها نصب است، ممکن است عیب بزرگی محسوب نشود.

ب. آنهایی که به نصب نرم‌افزار نیازی ندارند و معمولاً، برنامه‌نویسان از قبل توابع و برنامه‌هایی نوشته و درون اینترنت منتشر کرده‌اند که می‌توان با بهره‌گیری از آنها، حتی اگر برنامهٔ مربوطه روی رایانه نصب نباشد، می‌توان متن را به رشته تبدیل کرد. حُسن و عیب این مورد دقیقاً عکس مورد بالاست. در این مورد، نیازی به نصب برنامه‌هایی مانند Microsoft Office روی رایانه نیست اما در عوض باید فایل‌های جانبی (معمولاً فایل‌های با پسوند dll) به برنامه نصب نرم‌افزار نمایه‌سازی پیوست کرد.

۲. مسائل رسم الخط فارسی در نمایه‌سازی

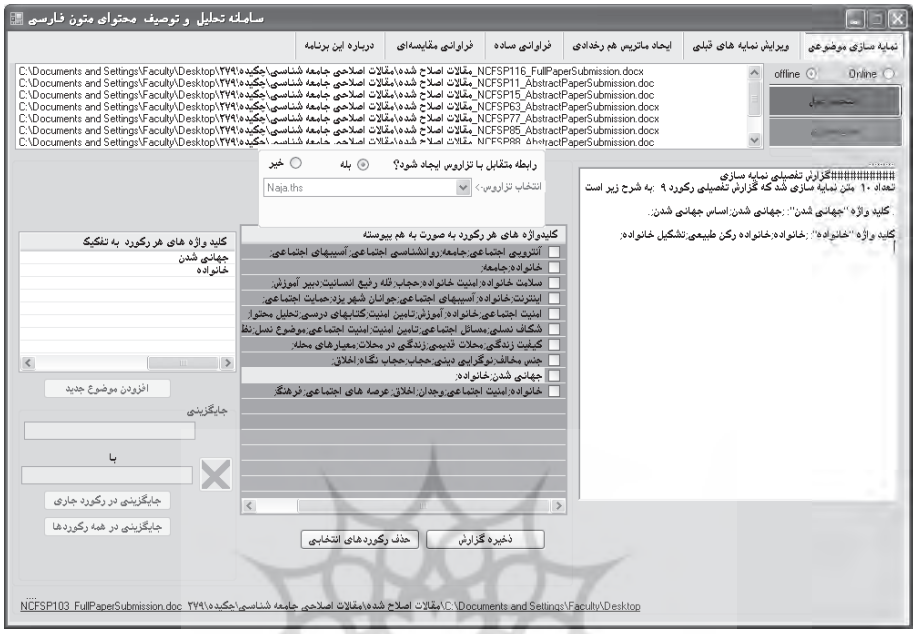
در رسم‌الخط فارسی به علت پیوسته بودن حروف به یکدیگر و به‌کارگیری استانداردهای گوناگون، مسائلی وجود دارد که باعث می‌شود متون از لحاظ نوشتاری یک‌دست نباشند. حاجی‌زین‌العابدینی(۱۳۷۸) در تحلیل و ارزیابی پایگاه اطلاعاتی کتابشناسی ملی و ارسطوپور و احمدی‌نسب(۱۳۹۰) درآسیب‌شناسی زبان و خط فارسی در بازیابی اطلاعات، به پاره‌ای از آنها اشاره کرده‌اندکه می‌تواند مرجع خوبی در این زمینه باشد.

مشکلات رسم‌الخط فارسی در نوشتن علامت‌های جمع، برخی پیشوندها و پسوندها ظاهر می‌شود. «می»، «ها» ، «تر» و «ترین» از شایع‌ترین آنها است. معمولاً گروه‌های سنی مختلف از رسم‌الخط گوناگون استفاده می‌کنند. برخی افراد هر یک از این پیشوندها و پسوندها را به واژهٔ اصلی می‌چسبانند، مثل «میشود»، «قشنگتر»، و «گنجشکها» و برخی دیگر، به‌صورت «می‌شود»، «قشنگ‌تر» و «گنجشک‌ها» می‌نویسند. پیچیده‌تر این‌که ممکن است هنگام تایپ، برای جدا کردن این پیشوند و پسوندها از فاصله یا نیم فاصله استفاده کنند. مثلاً در «می شود» بین «می» و «شود» یک فاصله وجود دارد در حالی که در «می‌شود» به‌دلیل استفاده از نیم‌فاصله بین آنها فاصله‌ای دیده نمی‌شود. بنابراین، باید پس از ریخته‌شدن متن روی رشته، ابتدا تکلیف این چیزها را مشخص کرد. مناسب‌ترین روش این است که ابتدا تمامی نیم‌فاصله‌ها را به فاصله تبدیل کرد. سپس هر جا که پیشوندهایی مثل «می» وجود دارد به واژه بعد از خود، و هر جا پسوندهایی مانند «ها» و چیزهای دیگر وجود دارد، به واژهٔ قبل از خود چسبانده شود.

مشکل دیگر گوناگونی استانداردهای به‌کار رفته در کدگذاری حروف است. در زمینهٔ استانداردها، دو حرف «یا» و «کاف» بسیار با اهمیت است. حرف یاء در عربی به صورت «ي» نوشته می‌شود که در برخی از سیتم عامل‌ها مانند Windows XP از "ی"، بدون دو نقطه در زیر، قابل تشخیص نیست. دیگری حرف «کاف» است که در عربی به شکل «ك» است در حالی که در فارسی به شکل «ک» با سرکش است. این مورد هم در سیتم عامل‌هایی مانند Windows XP قابل تشخیص نیست. افراد هنگام تایپ هر یک از متون فارسی ممکن است از یکی از این حالت‌ها بهره ببرند که سبب عدم یک‌دستی می‌شود. بنابراین، لازم است ابتدا به برنامه بگوییم که هر جا «ی» عربی است به «ی» فارسی و هر جا «کاف» عربی است به «کاف» فارسی یا بر عکس تبدیل شود.

سنجش توانمندی مدل

این مدل ابتدا برای نمایه‌سازی مقالات همایش‌ها طراحی گردید. هدف این بود که مقالات همایش‌ها به صورت یک‌دست نمایه‌سازی شوند و سپس نقشهٔ موضوعی همایش‌ها با فنون خوشه‌بندی و تحلیل شبکه‌های اجتماعی ترسیم گردد. برای این کار، در محیط ویژوال استودیو با زبان برنامه‌نویسی سی‌شارپ، مدل حاضر پیاده‌سازی شد که ظاهر این برنامه به شکل زیر بود:



تصویر ۱. ظاهر برنامهٔ طراحی شده بر اساس مدل دو مرحله‌ای شکاف-گلچین

تمامی چکیده‌های پذیرفته‌شده در همایش ملی خانواده و امنیت که تعداد آنها از ۵۰۰ مورد تجاوز می‌کرد، به‌صورت پایلوت با برنامهٔ یاد شده نمایه‌سازی شدند. از طرفی کلیدواژه‌های نویسندگان این مقالات مورد بازنگری قرار گرفت زیرا به علت ناآگاهی بیشتر نویسندگان از اصول علمی نمایه‌سازی اشکالاتی وجود داشت؛ از جمله این که مشکل افزونگی در بسیاری از کلیدواژه‌ها رخ داده بود. مشکل دیگر، به‌کارگیری واژه‌ها و عباراتی بود که از نگاه نمایه‌سازی، اصطلاح محسوب نمی‌شوند مانند «نگرش»، «تبیین» و غیره.

برای سنجش اعتبار و توانمندی یک مدل نمایه‌سازی، روش‌های گوناگونی از جمله روش کَتز[1] (۱۹۸۸) وجود دارد که سه شاخص برای سنجش اعتبار در نظر می‌گیرد: نرخ‌های از دست دادن پایداری[2]، خطای مثبت[3] و خطای منفی[4]. روش‌های دیگری نیز وجود دارد که به همخوانی یا شباهت‌سنجی نتایج نمایه‌سازی دو یا چند عامل تمرکز می‌کند. به نظر تونتا (۱۹۹۱)[5] پیش‌فرض این روش‌ها آن است که رابطه‌ای بین همخوانی در نمایه‌سازی و کیفیت نمایه‌سازی

1. Katz
2. fixation loss
3. false-positive
4. false-negative
5. Tonta

وجود دارد. در برابر مدلیان[1] و ویتن[2] (۲۰۰۶) معتقدند که اِشکال این روش‌ها نپرداختن به ارتباطات معنایی بین اصطلاحات است. یکی از این روش‌ها، روش سنتی هوپر (۱۹۶۵)[3] است که به روش جاکاردی نیز شناخته می‌شود و مبتنی بر فرمول زیر است:

$$CP(\%) = \frac{100A}{A + M + N}$$

در فرمول بالا، CP برابر با درصد همخوانی و A، M و N نیز به ترتیب نشانگر تعداد اصطلاحات مشترک، تعداد اصطلاحاتی که نمایه‌ساز M تشخیص داده ولی نمایه‌ساز N تشخیص نداده و تعداد اصطلاحاتی که نمایه‌ساز N تشخیص داده ولی نمایه‌ساز M به‌عنوان اصطلاح تشخیص نداده است.

فرمول هوپر در صورتی مؤثر است که تعداد اصطلاحات داده شده توسط نویسندهٔ مقاله و سامانهٔ نمایه‌سازی یکسان باشد که در عمل این‌گونه نیست. برای حل این مسأله می‌توان از نسخهٔ دیگر فرمول جاکاردی استفاده کرد که «شاخص در بردارندگی[4]» نامیده می‌شود (کین[5]، ۲۰۰۰) و در اثر سلتون و مک‌گیل[6] (۱۹۸۳) آمده است. اگر بخواهیم همانند فرمول قبل بیان کنیم، شاخص در بردارندگی چنین خواهد بود:

$$Inc(\%) = \frac{100A}{M}$$

برای سنجش مدل حاضر، همانند کین (۲۰۰۰) از شاخص دربردارندگی بهره گرفته شد زیرا این شاخص، سنجش را به صورت نامتقارن یا یک‌طرفه ممکن می‌سازد. در حالت نامتقارن[7]، شباهت نتیجهٔ «مدل نمایه‌سازی حاضر» با «کلیدواژه‌های نویسنده» از شباهت «کلیدواژه‌های نویسنده» با نتیجهٔ «مدل نمایه‌سازی حاضر» متفاوت است. به عبارتی ما می‌خواهیم بدانیم که عملکرد مدل ما چه میزان با کار نویسندهٔ مقاله برابر است و به دنبال این نیستیم که کار نویسنده چه میزان با عملکرد مدل ما شباهت دارد. به زبان دیگر، آنچه مهم است، این است که نتیجهٔ این مدل چه میزان از کلیدواژه‌های نویسنده را در بر دارد. از این رو، اگر در آزمودن مدل حاضر، کلیدواژه‌هایی علاوه بر کلیدواژه‌های نویسنده نتیجه بدهد، تأثیر منفی بر عملکرد مدل نخواهد داشت.

1. Medelyan
2. Witten
3. Hooper
4. Inclusion index
5. Qin
6. McGill & Salton
7. asymmetric

با استفاده از «شاخص دربردارندگی» شباهت نتایج حاصل از نمایه‌سازی با نرم‌افزار حاضر و کلیدواژه‌های نویسندگان سنجیده شد. اما قبل از آن وضعیت پروزن‌ترین اصطلاح هر مقاله مورد توجه قرار گرفت. در سامانه نمایه‌سازی که بر اساس مدل یاد شده طراحی شده است، همیشه پروزن‌ترین اصطلاح نمایه‌شده در ابتدا قرار می‌گیرد. از این جهت بررسی شد تا ببینیم که آیا حتماً این اصطلاح در بین کلیدواژه‌های نویسنده یا نویسندگان آن مقاله وجود دارد یا خیر. در نزدیک به ۹۵٪ از موارد جواب بله بود. یعنی هر اصطلاحی که سامانه نمایه‌سازی به‌عنوان پروزن‌ترین تشخیص داده بود، حتماً توسط نویسندگان نیز به عنوان کلیدواژه برگزیده شده بود و جالب‌تر از همه این که در نزدیک به ۹۰٪ از موارد اولین اصطلاح نمایه‌ای همان اولین کلیدواژه‌ای بود که نویسنده به مقاله خود داده بود. باید خاطرنشان ساخت که اگر اختلافات جزئی وجود داشت، اصطلاح نمایه‌ای و کلیدواژه شبیه دانسته می‌شد. مثلاً، اصطلاح نمایه‌ای «مصرف سیگار» و کلیدواژه «سیگار» شبیه در نظر گرفته شدند.

پس از آن از شاخص دربردارندگی استفاده شد. همان‌گونه که قبلاً ذکر شد، این شاخص شباهت را به‌صورت نامتقارن یا یک‌طرفه می‌سنجد یعنی به‌صورت یک طرفه، شباهت یا همخوانی عملکرد سامانهٔ نمایه‌سازی خودکار با کلیدواژه‌های نویسندگان سنجیده شد زیرا هدف این است که ببینیم اصطلاحات حاصل از سامانهٔ خودکار چه میزان از کلیدواژه‌های نویسندگان را در بر دارد. برای این کار، شباهت اصطلاحات نمایه‌ای هر مقاله با کلیدواژه‌های آن مقاله جداجدا سنجیده شد و سپس این مقدار به تعداد مقالات مورد بررسی تقسیم شد که فرمول ریاضی آن به شکل زیر است:

$$C = \frac{\sum \frac{\sum_d ai}{\sum_d a}}{D} \times 100$$

در فرمول بالا، C برابر با همخوانی، d برابر با چندمین مدرک، a برابر با کلیدواژه‌های نویسنده، i برابر با اصطلاحات نمایه‌شده و D تعداد کل مدارک است. بر این اساس، $\sum_d ai$ معادل تعداد موارد مشابه یا اشتراک مجموعه کلیدواژه‌های نویسنده و اصطلاحات نمایه‌ای است و $\sum_d a$ نیز تعداد کلیدواژه‌های نویسنده محسوب می‌شود.

تعداد مقالات نمایه‌شده توسط سامانهٔ خودکار، ۵۷۵ مورد بود. بنابراین، D برابر با این مقدار می‌شود. صورت کسر نیز برابر با ۴۳۰ بود که نتیجهٔ نهایی نشان داد که تقریباً ۷۶ درصد شباهت یا همخوانی بین سامانهٔ خودکار و کلیدواژه‌های نویسندگان وجود دارد.

نتیجه‌گیری

در این نوشتار، ابتدا مدل دو مرحله‌ای شکاف-گلچین معرفی شد. اَلگوریتم کلی پیاده‌سازی این مدل را می‌توان به صورت زیر خلاصه کرد:

- **مقدمات**
  - فراخوانی فایل متن به برنامه و تبدیل متن به رشته؛
  - یک‌دست‌سازی رسم‌الخط متن؛
- **شکاف‌دهی**
  - تبدیل علائم به شکاف؛
  - تبدیل اَفعال به شکاف؛
  - تبدیل واژه‌ها و عبارات ممنوعهٔ نوع ۱ به شکاف؛
  - تبدیل واژه‌ها و عبارات ممنوعهٔ نوع ۴ به شکاف (البته آنهایی که ممنوعه هستند)؛
  - تبدیل واژه‌ها و عبارات ممنوعهٔ نوع ۲ به شکاف؛
  - تبدیل واژه‌ها و عبارات ممنوعهٔ نوع ۳ به شکاف؛
- **گلچین کردن**
  - تصمیم‌گیری دربارهٔ ارجحیت حالت جمع یا مفرد واژه‌ها،
  - وزن‌دهی اصطلاحات بین‌شکافی؛
  - کنار گذاشتن اصطلاحات افزونه (تهیهٔ فهرست اولیهٔ اصطلاحات گلچین شده)؛
  - تهیهٔ فهرست نهایی اصطلاحات گلچین شده.

همان‌گونه که از تصویر ۱ بر می‌آید، این مدل به دو صورت خودکار و نیمه‌خودکار قابل پیاده‌سازی است. در این مدل گر چه برای شکاف دادن، به اجبار بخش‌هایی از متن حذف می‌شود اما در نهایت اصطلاحات تشخیص داده شده توسط مدل، نزدیکی بسیاری با کلیدواژه‌های نویسندگان دارد. ارزیابی مدل حاضر، همانند سایر ارزیابی‌هایی که از مدل‌های نمایه‌سازی می‌شود، به مجموعه‌ای کوچک (پایلوت) از مدارک محدود شده است. علاوه بر آن، نمایه‌سازی فقط روی چکیده‌ها صورت گرفته است. با این وصف انتظار می‌رود که نمایه‌سازیِ متون طولانی مانند کتاب‌ها و متن کامل مقالات نیز با این روش ممکن باشد. از

مزیت‌های دیگر این مدل آن است که عمق لازم برای نمایه‌سازی را تشخیص می‌دهد. عمق نمایه‌سازی که ترکیبی از جامعیت و مانعیت است، با تعداد اصطلاحات گلچین‌شده برای یک مدرک مرتبط است. تابع تعیین تعداد اصطلاحات انتخابی در این مدل، بدون توجه به طول متن، تعداد اصطلاحات مورد نیاز و لازم برای توصیف یک مدرک را مشخص می‌سازد.

بدیهی است که نمی‌توان سامانه‌ای در حد ایده‌آل ایجاد کرد. باید همانند سایر محصولات فنی، سامانه‌های گوناگونی طراحی شود و با درک ضعف‌ها و قوت‌های عملی آن، به بهبود آنها اقدام کرد. مزیت این مدل آن است که حداقل از حد نظریه خارج شده است و به عمل نشسته است و برای بهبود و توسعه، قابل نقد است.

کین (۲۰۰۰) در مقایسه‌ای که روی پایگاه استنادی علوم[1] و مدلاین[2] انجام داد، دریافت که اصطلاحات عام مدارک این دو پایگاه با هم شباهت دارند و این اصطلاحات خاص هستند که عدم همخوانی را به وجود می‌آورند. این نکته در نتایج حاصل مورد توجه قرار گرفت و مشاهده شد که سامانهٔ خودکار و کلیدواژه‌های نویسندگان در تشخیص اصطلاحات خاص عدم همخوانی کامل دارند. در بخش یافته‌ها ذکر شد که بین پروزن‌ترین اصطلاح به‌دست آمده در هر مقاله با اولین کلیدواژهٔ نویسندگان نزدیک به ۹۰ درصد تشابه وجود دارد. معمولاً نویسندگان عام‌ترین موضوعات را به‌عنوان اولین کلیدواژه‌های خود می‌نویسند و این اصطلاحات عام در متن بیشتر تکرار می‌شوند. از این رو، سامانهٔ خودکار در تشخیص این نوع از اصطلاحاتِ موضوعی موفق بوده است و برای تشخیص اصطلاحات خاص باید پارامترهای دیگری غیر از پارامترهای چهارگانهٔ این مدل برای وزن‌دهی به اصطلاحات در نظر گرفت. بارتو[3] (۲۰۱۲) در پژوهش خود دریافت که امروزه در نمایه‌سازی گرایش به سمت انتخاب اصطلاحات خاص است. بنابراین، در مدل‌های نمایه‌سازی باید به تشخیص اصطلاحات اخص توجه داشت.

از اولین پژوهش‌هایی که در زمینهٔ همخوانی نمایه‌سازی انجام شده است می‌توان به اثر لنکستر[4] (۱۹۶۸) اشاره کرد. وی با استفاده از فرمول هوپر دریافت که همخوانی بین دو نمایه‌ساز (عامل انسانی) برای نمایه‌سازی مقالات پزشکی در پایگاه مدلاین بین ۳۴/۴ درصد تا ۴۶/۱ درصد است. لئونارد[5] (۱۹۷۵) پژوهش دیگری در همین زمینه انجام داد و نتایجی نزدیک به کار

1. Science Citation Index
2. MEDLINE
3. Barto
4. Lancaster
5. Leonard

لنکستر یعنی ۳۶/۵ درصد تا ۴۸/۲ درصد به دست آورد. در دهه‌های اخیر، نتایج مشابهی به‌دست آمده است. مثلاً لایننگر[1] (۲۰۰۰) با روش هوپر میانگین این مقدار را ۵۰/۴ درصد به دست آورد. با توجه به بررسی‌های قبلی، نتیجهٔ به‌دست آمده از طریق نمایه‌سازی با مدل پیشنهادی را می‌توان در حد قابل قبول دانست.

منابع

ارسطوپور، شعله و احمدی‌نسب، فاطمه (۱۳۹۰). آسیب‌شناسی زبان و خط فارسی در بازیابی اطلاعات: نگاهی به موتورهای کاوش و پایگاه‌های برخط. *مجموعه مقالات نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب*. تهران: کتابخانهٔ ملی جمهوری اسلامی ایران.

برنجیان، شاپور رضا (۱۳۹۰). *ریشه‌یاب ماضی و مضارع از مصدر اَفعال ناگذر در زبان فارسی*. شیراز : نوید شیراز؛ مرکز منطقه ای اطلاع‌رسانی علوم و فن‌آوری.

حاجی‌زین‌العابدینی، محسن (۱۳۷۸). تحلیل و ارزیابی پایگاه اطلاعاتی کتابشناسی ملی ایران. در: فهرست‌های رایانه‌ای: کاربرد و توسعه، *مجموعه مقالات همایش کاربرد و توسعهٔ فهرست‌های رایانه‌ای در کتابخانه‌های ایران*. مشهد: دانشگاه فردوسی مشهد؛ تهران: مرکز اطلاع‌رسانی و خدمات علمی جهاد دانشگاهی.

دانشکار آراسته، پویا (۱۳۸۳). نرم‌افزار تشخیص فعل در زبان فارسی. *مجلهٔ فرهنگ*، ۴۹ و ۵۰، ۳۱-۴۶.

مولودی، امیرسعید (۱۳۹۰). *فعل مرکب و معیارهای صوری برای تشخیص آن*. دبیرخانهٔ شورای عالی اطلاع‌رسانی، مرجع دادگان زبان فارسی. بازیابی شده در تاریخ ۳۰ مهرماه ۱۳۹۲ از:
http://dadegan.ir/content/%D9%81%D8%B9%D9%84-%D9%85%D8%B1%DA%A9%D8%A8-%D9%88-%D9%85%D8%B9%DB%8C%D8%A7%D8%B1%D9%87%D8%A7%DB%8C-%D8%B5%D9%88%D8%B1%DB%8C-%D8%A8%D8%B1%D8%A7%DB%8C-%D8%AA%D8%B4%D8%AE%DB%8C%D8%B5-%D8%A2%D9%86

References

Anderson, J. D. & Pérez-Carballo, J. (2001). 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 & Management*, 37 (2), 231-254.

Barto, T. (2012). Assessment of Indexing Trends with Specific and General Terms for Herbal Medicine Health. *Information & Libraries Journal*, 29 (4), 285–295.

Giyeong, K. (2006). Relationship between Index Term Specificity and Relevance Judgment. *Information Processing and Management: an International Journal*, 42 (5), 1218 – 1229.

1. Leininger

Hooper, R. S. (1965). *Indexer Consistency Tests: Origin, Measurement, Results, and Utilization*. Bethesda, Maryland: IBM Corporation.

Jenuwine, E. S. & Floyd, J. A. (2004). Comparison of Medical Subject Headings and Text-Word Searches in MEDLINE to Retrieve Studies on Sleep in Healthy Individuals. *Journal of Medicine Library Association*, 92 (3), 349-354.

Jones, K. S. (2004). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 60 (5), 493-502.

Kageura, K. & Umino, B. (1996). Methods of Automatic Term Recognition: A review. *Terminology*, *3* (2), 259-289.

Katz, J. (1988). A Reliability Indexes of Automated Perimetric Tests. *Archives of Ophthalmology*, 106 (9), 1252- 1254.

Lancaster, F. W. (1968). *Evaluation of the MEDLARS Demand Search Service*. Washington, D.C.: Na-Library of Medicine.

Leininger, K. (2000). Interindexer Consistency in PsycINFO. *Journal of Librarianship and Information Science*, 32 (1), 4-8.

Leonard, L. E. (1975). *Inter-Indexer Consistency and Retrieval Effectiveness: Measurement of Relationships*, PhD Thesis, Illinois: University of Illinois.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2 (2), 159-165.

Mai, J. E. (2001). Semiotics and Indexing: an Analysis of the Subject Indexing Process. *Journal of Documentation*, 57 (5), 591-622.

Maron, M. E. (1979). Depth of Indexing. *Journal of the American Society of Information Science*, 30 (4), 224-228.

Medelyan, O. & Witten, H. L. (2006). Measuring Inter-Indexer Consistency Using a Thesaurus**.** *Proceedings of JCDL '06 Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries,* June 11 - 15, *pp.* 274-275, New York, USA.

Park, H. (1996). Inferential Representation of Science Documents. *Information Processing & Management*, 32 (4), 419-429.

Qin, J. (2000). Semantic Similarities between a Keyword Database and a Controlled Vocabulary Database: An Investigation in the Antibiotic Resistance Literature. *Journal of the American Society for Information Science*, 51 (2), 166-180.

Lancaster, F. W. (1991). *Indexing and Abstracting in Theory and Practice*. Champaign, IL: University of Illinois.

Raghavan, V. V. et al. (2004). *Information Retrieval. In the Practical Handbook of Internet Computing*, (Munindar P. Singh, ed.), Part-2, Chapter 12, Boca Raton, Florida: Chapman and Hall; CRC Press.

Salton, G.; Buckley, C. (1988). Term-Weighing Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24 (5), 513-523.

Taghva, K.; Beckley, R. & Sadeh, M. (2003). *A List of Farsi Stopwords*. Technical Report 2003-01, Information Science Research Institute, Las Vegas: University of Nevada.

Tonta, Y. (1991). A Study of Indexing Consistency Between Library of Congress and British Library Catalogers. *Library Resources & Technical Services*, 35 (2), 177-185.

Salton, G. & McGill, M. J.(1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Soergel, D. (1994). Indexing and Retrieval Performance: The logical evidence. *Journal of the American Society of Information Science*, 45 (8), 589-599.

Vu, T.; Aw, A. T. & Zhang, M. (2008). Term Extraction through Unit hood and Term hood Unification. *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, January, 7-12, Hyderabad, India.

به این مقاله این‌گونه استناد کنید:

توکلی‌زاده راوری، محمد (۱۳۹۴). مدل دو مرحله‌ای شکاف-گلچین برای نمایه‌سازی خودکار متون فارسی. *تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی*، ۲۱ (۱)، ۱۳-۴۰.