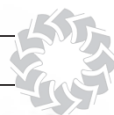


گسترش‌پذیری جستجو و بازیابی مدارک در پایگاه‌های اطلاعات علمی فارسی: مورد پژوهشی پیوسته‌نویسی و جدانویسی

ایوب رنجبر^۱، جواد عباس‌پور^۲



تاریخ ارسال: ۹۷/۴/۲۲ - تاریخ پذیرش: ۹۷/۵/۲۹

چکیده

هدف: هدف این مقاله مطالعه میزان توجه سه پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور»، «مرکز اطلاعات علمی جهاد دانشگاهی» و «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» به گسترش‌پذیری جستجو و بازیابی مدارک (مقالات) به صورت پیوسته‌نویسی و جدانویسی است.

روش‌شناسی: پژوهش حاضر از نظر هدف در ردیف پژوهش‌های کاربردی و از نظر طرح پژوهش از نوع تحلیل محتوا بوده است. جامعه آماری شامل مجموعه مدارک موجود در سه پایگاه اطلاعات علمی فارسی بود (۲۹۱۴۲ مدرک) که با استفاده از فرمول کوکران، ۴۳۲ مدرک از آنها با روش نمونه‌گیری طبقه‌ای متناسب (نسبتی) انتخاب شد. ابزار پژوهش یک سیاهه واری محقق‌ساخته بود. سیاهه واری شامل ده قاعده از مجموعه قاعده‌های پیوسته‌نویسی و جدانویسی کتاب «دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی» به همراه مصداق‌های «واژگان یا کلیدواژه‌های انتخابی» از هر قاعده بود که امکان رخداد جدانویسی و پیوسته‌نویسی آنها در هنگام نگارش وجود داشت.

۱. دانش‌آموخته علم اطلاعات و دانش‌شناسی، دانشگاه شیراز (نویسنده مسئول).

ayoubbranjar69@gmail.com

۲. عضو هیئت علمی گروه علم اطلاعات و دانش‌شناسی دانشگاه شیراز.

Javad.abbaspour@gmail.com

یافته‌ها: نتایج نشان داد میان پایگاه‌های اطلاعات علمی فارسی از نظر قابلیت بازیابی مدارک با هر یک از شکل‌های نگارشی اعم از پیوسته‌نویسی و جدانویسی تفاوت وجود دارد. علاوه بر این، یافته‌ها نشان داد، تنها ۳/۵٪ از مدارک انتخابی در هنگام جستجو در پایگاه‌های اطلاعاتی مورد مطالعه با همه شکل‌های نگارشی قابل بازیابی بود. در مقابل، ۹۴/۷٪ مدارک انتخابی فقط با همان حالت ثبت‌شده در پایگاه بازیابی شد. همچنین، مقایسه شکل نگارشی مصداق‌ها در عنوان، چکیده و کلمات کلیدی نسخه پی.دی.اف. مدارک با اطلاعات نمایه‌شده از همان مدارک در پایگاه‌های اطلاعات علمی فارسی نشان داد، شکل نگارشی بخش قابل توجهی از مصداق‌های مدارک از حالت نگارشی نزدیک‌نویسی به جدانویسی یا پیوسته‌نویسی تغییر یافته است.

نتیجه‌گیری: با توجه به اینکه در پایگاه‌های مورد مطالعه چالش‌های پیوسته‌نویسی و جدانویسی به طور جامع و به منظور بهبود جامعیت نتایج جستجو مورد توجه قرار نگرفته است، این وضعیت نامطلوب می‌تواند به ازدست‌دادن مدارکی بینجامد که با دیگر شکل‌های نگارشی در پایگاه‌های اطلاعاتی ذخیره شده‌اند.

کلیدواژه‌ها: بازیابی اطلاعات، پایگاه‌های اطلاعات علمی، پیوسته‌نویسی، جدانویسی، خط فارسی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

مقدمه

از مهم‌ترین مسائل و چالش‌های نگارشی خط فارسی در محیط‌های الکترونیکی (روند ذخیره‌سازی و پردازش، جستجو و بازیابی اطلاعات) می‌توان به مواردی چون نگارش غیررسمی یا محاوره‌ای، استفاده از واژگان بیگانه، پیچیدگی صرفی، تنوع املایی واژگان، هم‌نگاشت‌ها^۱ و فاصله‌گذاری میان واژگان اشاره کرد (دانesh، مینایی و کاشفی^۲، ۲۰۱۱). در میان این چالش‌ها، در دهه‌های اخیر مسئله «فاصله‌گذاری» و به طور مشخص پیوسته‌نویسی و جدانویسی^۳ بسیار مورد توجه قرار گرفته است؛ به گونه‌ای که

1. Homographs

2. Danesh, Minaei and Kashefi

۳. در عمده کتاب‌های دستور زبان فارسی از جمله کتاب «دستورخط فارسی مصوب فرهنگستان زبان و ادب فارسی» منظور از «جدانویسی» حالت نیم‌فاصله‌نویسی یا نزدیک‌نویسی است؛ این در حالی است که در این مقاله منظور از جدانویسی استفاده از فاصله کامل بین اجزای یک واژه واحد است و در جایی نیز که منظور حالت نیم‌فاصله بوده، از واژه «نزدیک‌نویسی» استفاده شده است.

بیشترین اختلاف نظرها در باب املائی فارسی (فرهنگستان زبان و ادب فارسی، ۱۳۸۹) و بیشترین پژوهش‌ها در ارتباط با چالش‌های نگارشی زبان فارسی در محمل‌های الکترونیکی (هنرجویان، ۱۳۹۲) را به خود اختصاص داده است. حتی، در کنار چالش‌های اعراب‌گذاری، از آن به عنوان یکی از مهم‌ترین و تأثیرگذارترین چالش‌ها در قلمرو منابع الکترونیکی یاد می‌شود، به گونه‌ای که صاحب نظران مراحل بعدی پردازش متون تولیدی را منوط به حل شدن این دو مسئله دانسته‌اند (کاشفی، نصری و کنعانی، ۱۳۸۹؛ دانش و دیگران، ۲۰۱۱).

به دلیل وجود این چالش در رسم الخط فارسی، اگر شکل‌های مختلف نگارشی یک واژه مرکب اعم از پیوسته نویسی یا جدانویسی توسط طراحان و نمایه‌سازان پایگاه‌های اطلاعات علمی فارسی بهنجارسازی^۱ نشود، کاربران با نگارش هر یک از شکل‌های نگارشی واژه، تنها به مدارکی^۲ که بدان شکل نگارش یافته است، دسترسی خواهند یافت و سایر مدارک مرتبط نگارش یافته با دیگر شکل‌ها را از دست خواهند داد.

برای تبیین بهتر مسئله و اینکه چه ناهماهنگی‌هایی در ضبط یک واژه واحد در محمل‌های الکترونیکی و به طور اخص پایگاه‌های اطلاعاتی علمی فارسی رخ می‌دهد، نگارنده از بخش «ترکیبات»^۳ دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی و از قاعده ده که مربوط به پیوسته نویسی است، واژه «کتابشناسی» را به عنوان نمونه انتخاب و در پایگاه استنادی علوم جهان اسلام^۴ (به عنوان نمونه‌ای از پایگاه‌های اطلاعات علمی فارسی) جستجو کرد. نگارش این واژه به سه شکل «کتابشناسی» (پیوسته نویسی)، «کتاب شناسی» (نزدیک نویسی) و «کتاب شناسی» (جدانویسی)

1. Normalization

۲. در این پژوهش منظور از «مدارک / مدارک»، مقاله یا مقالات علمی - پژوهشی و علمی - ترویجی نمایه شده در پایگاه‌های اطلاعات علمی فارسی مورد مطالعه است.

۳. در کتاب دستور خط فارسی مصوب فرهنگستان «ترکیبات» شامل واژگان مرکب و مشتق است.

4. ISC (Islamic World Science Citation Center)

امکان پذیر است. در ابتدا جستجو به شکل «کتاب شناسی» (جدانویسی) انجام شد که ۳۶۲ مدرک بازیابی شد؛ ولی با تغییر شکل نگارشی واژه و وارد کردن آن به شکل «کتاب شناسی» (نزدیک نویسی) و «کتابشناسی» (پیوسته نویسی) به ترتیب ۴۰ و ۵۴ مدرک بازیابی شد. توجه نکردن به چنین مسئله‌ای در مراحل ذخیره‌سازی و پردازش، جستجو و بازیابی اطلاعات از پایگاه‌های اطلاعات علمی فارسی، چالش‌ها و موانعی را پیش روی کاربران فارسی‌زبان در دستیابی به اطلاعات مورد نیازشان ایجاد کرده است.

مرور مطالعات انجام شده در سال‌های اخیر نشان می‌دهد میزان توجه سامانه‌های بازیابی اطلاعات به چالش‌های خط فارسی، از جمله مسئله پیوسته‌نویسی و جدانویسی، بعضاً با ذکر تعداد محدودی کلیدواژه (یک یا دو کلیدواژه) آن‌هم از قلمرو موضوعی محدود، به منزله نمونه‌ای از متون فارسی، بررسی شده است. افزون بر این، یکسانی فراوانی تعداد مدارک بازیابی شده در ریخت‌های مختلف یک واژه را نشان از رفع آن چالش در سامانه‌های بازیابی اطلاعات قلمداد کرده‌اند. این در حالی است که ممکن است شکل‌های نگارشی یک واژه به‌طور همزمان در بخش عنوان، چکیده و واژگان کلیدی یک مدرک رخ داده و این مسئله باعث بازیابی دوباره همان مدرک شده باشد.

علاوه بر موارد بالا، در بخش ترکیبات دستور خط فارسی فرهنگستان، ۲۳ قاعده درباره موقعیت‌های رخداد جدانویسی و پیوسته‌نویسی ذکر شده است. این در حالی است که مرور مطالعات پیشین (رمضانی، ۱۳۸۶؛ قدس‌نیا، زارع بیدکی، و یزدانی، ۱۳۸۶؛ عبدالهی و جوکار، ۱۳۸۸؛ گل تاجی و بذرگر، ۱۳۸۹؛ آخشیک و فتاحی، ۱۳۹۱؛ هماوندی، نوروزی و حسینی بهشتی، ۱۳۹۷) نشان می‌دهد پژوهشگران بدون توجه به این قاعده‌ها، گاه با در نظر گرفتن یک یا دو مصداق (کلیدواژه یا واژگان انتخابی) آن‌هم بدون ذکر چگونگی و دلایل انتخاب آنها، چالش‌های پیوسته‌نویسی و جدانویسی و همچنین میزان تأثیر آن بر سامانه‌های بازیابی اطلاعات را بررسی کرده‌اند. در صورتی که به نظر می‌رسد انتخاب مصداق (کلیدواژه) از هر قاعده می‌تواند تأثیرگذاری متفاوتی بر

سامانه‌های جستجو و بازیابی اطلاعات داشته باشد.

با توجه به وجود چنین خلأهایی در پژوهش‌های پیشین، مقاله حاضر در پی آن است که با گسترش قلمروهای موضوعی و با در نظر گرفتن قاعده‌های بخش ترکیبات دستور خط فارسی مصوب فرهنگستان و همچنین، کنترل رخداد شکل‌های نگارشی در بخش عنوان، چکیده و واژگان کلیدی هر مدرک، وضعیت برخی از پایگاه‌های اطلاعات علمی فارسی را از نظر میزان توجه به مسئله پیوسته‌نویسی و جدانویسی بررسی و بر مبنای نتایج حاصل، پیشنهادها و راهکارهای عملی را ارائه کند. انتظار می‌رود با استفاده از نتایج این پژوهش بتوان از طریق تشخیص دقیق‌تر نقاط ضعف و قوت پایگاه‌های اطلاعات علمی فارسی در راستای بهبود یا اصلاح الگوریتم‌های جستجو و بازیابی آنها گام برداشت. همچنین، نمایه‌سازان و طراحان سامانه‌های بازیابی اطلاعات می‌توانند با شناسایی قاعده‌های دارای تأثیرگذاری بیشتر بر کمیّت نتایج بازیابی شده، امکان به‌کارگیری و اعمال آنها در الگوریتم‌های نمایه‌سازی یا بازیابی را فراهم کنند. از این رو، مقاله حاضر در پی پاسخ به سؤال‌های زیر نگارش یافته است:

۱. آیا میان پایگاه‌های اطلاعات علمی فارسی از نظر گسترش‌پذیری جستجو و بازیابی مدارک به صورت پیوسته‌نویسی و جدانویسی، تفاوت وجود دارد؟
۲. نسبت فراوانی پایگاه‌های اطلاعات علمی فارسی از نظر گسترش‌پذیری جستجو و بازیابی مدارک به صورت پیوسته‌نویسی و جدانویسی بر حسب قاعده‌های دهگانه فرهنگستان زبان و ادب فارسی چگونه است؟
۳. نسبت فراوانی شکل نگارشی مصداق‌ها در عنوان، چکیده و واژگان کلیدی نسخه‌پی‌دی‌اف مدارک با اطلاعات ارائه شده از همان مدارک در پایگاه‌های اطلاعاتی علمی فارسی، چگونه است؟

شایان ذکر است، در مقاله حاضر منظور از گسترش‌پذیری جستجو، توانایی پایگاه اطلاعات علمی فارسی در بازیابی مدرک با هر دو حالت نگارشی مصداق اعم از

پیوسته‌نویسی و جدانویسی است. بنابراین، سنجش گسترش‌پذیری جستجو برای هر پایگاه اطلاعات علمی فارسی با مشاهده رویداد یک یا مجموعه‌ای از حالت‌های زیر در زمان جستجوی مصداق‌های مدارک تعیین گردید:

۱. خوانش شکل نگارشی «پیوسته‌نویسی» و گسترش به حالت «جدانویسی»؛
۲. خوانش شکل نگارشی «جدانویسی» و گسترش به حالت «پیوسته‌نویسی»؛
۳. خوانش شکل نگارشی «جدانویسی» و گسترش به حالت «نیم‌فاصله»؛
۴. خوانش شکل نگارشی «پیوسته‌نویسی» و گسترش به حالت «نیم‌فاصله»؛
۵. خوانش «نیم‌فاصله» و گسترش به هر سه حالت نگارشی (پیوسته، نیم‌فاصله و جدا).

مرور پیشینه‌های پژوهش

پیشینه‌های داخلی

تاکنون ویژگی‌ها و مشکلات رسم الخط یا شیوه خط فارسی، از جمله مسئله فاصله‌گذاری و تأثیرگذاری آنها بر روند سامانه‌های ذخیره و بازیابی اطلاعات، موضوع پژوهش‌های متعددی بوده است. «حزّی» (۱۳۷۲) در مقاله‌ای نظری به تبیین چالش‌های رسم الخط فارسی در مواجهه با رایانه پرداخت. او راه‌حل‌های طراحی شده برای مقابله با چالش‌های رسم الخط را در پنج دسته کلی، یعنی ۱. هماهنگ‌کردن حروف ۲. استفاده از تکواژها ۳. استفاده از سیاهه آماده ۴. پیوند ساختگی میان واژگان و ۵. هماهنگی رسم الخط جای داد. او همچنین از میان راهکارهای پیشنهادی، مورد پنجم را ارجح و معقول‌تر دانست. حزّی دلیل آن را چنین تبیین می‌کند که با به‌کارگیری این راهکار، به دلیل ضرورت برخورد با رایانه، خط فارسی شیوه‌ای واحد خواهد یافت و آشفتگی و چندگانگی فعلی رسم الخط از بین خواهد رفت.

«اکبری نژاد» (۱۳۷۶) به واسطه تجربه کار با پایگاه‌های اطلاعاتی، در مقاله‌ای با عنوان «فاصله خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات» به بیان مشکلات و

مسائل ایجادشده به دلیل فاصله میان واژه‌ها و عبارات در تمام نظام‌های رایانه‌ای ذخیره و بازیابی اطلاعات کتاب‌شناختی به زبان فارسی پرداخت. نگارنده با اشاره به اینکه ملاک شناسایی واژه‌ها و نمایه‌سازی در نرم‌افزارهای موجود بازار، فضای خالی میان واژه‌هاست، از مسائل و مشکلات موجود در رسم‌الخط از جمله فاصله‌گذاری میان واژه‌ها و عبارات به صورت فاصله، نیم‌فاصله یا بی‌فاصله، به عنوان عاملی تأثیرگذار (منفی) بر جامعیت مطلوب یا سرعت جستجو یاد کرده است.

«مرتضایی» (۱۳۸۰) نیز با هدف ارائه نمونه‌هایی از تجربه‌های واژه‌گزینی در ذخیره اطلاعات و به منظور تسریع و تسهیل ذخیره و بازیابی اطلاعات، به بررسی مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات پرداخت. یافته‌ها نشان داد مسائل زبان و خط فارسی سبب کندی مراحل ذخیره و بازیابی اطلاعات، کاهش نسبت بازیافت اطلاعات و همچنین تأثیر منفی بر جامعیت نتایج یک جستجو می‌شود.

از جمله پژوهش‌های دیگر می‌توان به پژوهش «راثی» (۱۳۸۵) اشاره کرد که با استفاده از پرسش‌نامه و با مطالعه موردی کاربران مرکز اینترنت دانشگاه آزاد اسلامی شبستر، مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت را بررسی کرد. پژوهش به روش پیمایشی و با رویکرد توصیفی صورت گرفت. نتایج نشان داد کاربران کمترین استفاده را از «شکل‌های مختلف نوشتاری واژه» دارند و بیشتر کاربران به این نکته توجهی نداشتند. یافته‌های پژوهش‌گر این فرضیه را که «بیشتر از ۵۰٪ موارد عدم بازیابی مطالب مورد نظر در جستجوی اطلاعات به زبان فارسی، مربوط به مسائل زبان و خط فارسی است» تأیید کرد.

«قدس‌نیا، زارع بیدکی و یزدانی» (۱۳۸۶) در پژوهشی دیگر که با هدف سنجش تأثیرگذاری سیزده چالش از مهم‌ترین مشکلات زبان و خط فارسی بر میزان جامعیت نتایج جستجو انجام گرفت، با طراحی یک خزنده^۱ یک میلیون و دویست هزار صفحه

وب را بررسی کردند. نتایج پژوهش آنان نشان داد جدی‌ترین چالش‌ها و مشکلاتی که بر جامعیت نتایج جستجو تأثیرگذار است، به ترتیب شامل استفاده از «ی» فارسی و «ی» عربی نقطه‌دار چسبان به جای «ی» عربی آخر با ۹۴/۹۴٪، استفاده از «آ» به جای «ا» با ۹۴/۴۴٪ و استفاده از «ی» عربی نقطه‌دار چسبان و «ی» عربی آخر به جای «ی» فارسی با ۷۴/۹۸٪ است؛ همچنین استفاده از «فاصله» به جای «نیم فاصله» با ۵۱/۷٪ و «نیم فاصله» به جای «فاصله» با ۴۸/۳٪ از دیگر موارد قابل توجه است.

«عبداللہی و جوکار» (۱۳۸۸) در مقاله‌ای به بررسی چالش‌های شیوہ نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش گوگل، یاهو و آلتا ویستا پرداختند. براین مبنا، پژوهشگران به روش پیمایشی - مقایسه‌ای و اسنادی و با استفاده از یک سیاهه شامل هفده کلیدواژه که هر یک نمایانگریک مورد از چالش‌های زبان فارسی در بازیابی اطلاعات بود، این مطالعه را انجام دادند. یافته‌های پژوهش نشان داد موتورهای کاوش وب، شیوہ‌های نگارش زبان فارسی را به منظور بهبود کاوش، مورد توجه قرار نداده‌اند و رابطه معناداری بین شکل واژه و نوع ابزار جستجو وجود دارد. در ارتباط با این پژوهش، نکته قابل تأمل اینکه پژوهشگران برای سنجش تأثیر پیوسته نویسی و جدانویسی بر بازیابی اطلاعات، تنها به دو مصداق بسنده کرده‌اند و از چگونگی و دلایل انتخاب این دو مصداق نیز سخنی به میان نیاورده‌اند. همچنین، پژوهشگران پس از بررسی انجام شده در زمینه این چالش، به یکسان بودن فراوانی مدارک بازیابی شده اشاره کردند و از این طریق درباره عملکرد موتورهای جستجو نسبت به رفع احتمالی چالش‌های رسم الخط فارسی، به نتیجه گیری پرداختند. در توضیح این مطلب باید گفت، یکسان بودن فراوانی نتایج بازیابی شده به معنای یکسان بودن رخداد همان مدارک نیست زیرا احتمال دارد شکل‌های مختلف یک واژه در بخش‌های مختلف یک مدرک (عنوان، چکیده و کلیدواژه‌ها) رخ داده و به بازیابی دوباره آن مدارک انجامیده باشد. از سوی دیگر، ممکن است فقط یک حالت نگارشی در مدرک رخ داده باشد و پایگاه اطلاعات علمی فارسی،

امکان بهنجارسازی شکل‌های مختلف واژه را داشته باشد.

«گل‌تاجی و بذرگر» (۱۳۸۹) مسائلی ریخت‌شناسی زبان فارسی را در سه پایگاه مقاله‌های فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»، «مرکز اطلاعات علمی جهاد دانشگاهی» و «پژوهشگاه اطلاعات و مدارک علمی ایران (به اختصار ایرانداک)»^۱ با استفاده از روش پیمایشی - مقایسه‌ای بررسی کردند. در این پژوهش، پژوهشگران از یک سیاهه شامل ۱۷ کلیدواژه استفاده کردند که به اعتقاد آنها هرکدام نمایانگر یک مورد از چالش‌های زبان فارسی در برخورد با فناوری نوین بود. نتایج کلی پژوهش آنان نشان داد هیچ‌کدام از سه پایگاه فارسی مورد بررسی، به شیوه‌ای جامع چالش‌های زبان‌شناختی زبان فارسی را در جهت بهبود نتایج جستجو مورد توجه قرار نداده‌اند. همچنین نگارندگان در بعضی موارد تساوی تعداد رکورد‌های بازیابی شده در ریخت‌های مختلف یک واژه را نشان از رفع آن چالش خاص در نظر گرفته‌اند. در ارتباط با پژوهش «گل‌تاجی و بذرگر» نیز موارد قابل تأملی که پیش از این درباره پژوهش «جوکار و عبداللهی» اشاره شد، مصداق می‌یابد.

در مقاله «آخشیک و فتاحی» (۱۳۹۱) با عنوان «تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی»، آنها صد عنوان از پایان‌نامه‌های موجود رشته کتابداری و اطلاع‌رسانی را به‌منزله نمونه‌ای از متون فارسی در پایگاه‌های اطلاعاتی «پژوهشگاه‌های علوم و فناوری اطلاعات ایران» و «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» انتخاب و تحلیل محتوا کردند. سپس، از هر پایگاه ۵۰ عنوان را به صورت تصادفی انتخاب و عنوان‌های مربوط به هر پایگاه را با «جستجوی عنوانی» به‌طور عمده در حالت‌های درست و نادرست جستجو کردند. نتایج پژوهش آنان نشان داد، هرچند در پایگاه اطلاعاتی پژوهشگاه علوم و فناوری

۱. نام پیشین این پایگاه مرکز اطلاعات و مدارک علمی ایران بوده است.

۲. علم اطلاعات و دانش‌شناسی

اطلاعات ایران ۵۸٪ عنوان‌ها با تغییر شکل نگارشی مجدد بازیابی می‌شدند؛ اما در پایگاه «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» فقط حالت ثبت شده عنوان‌ها به بازیابی مجدد آن عنوان می‌انجامید. در ارتباط با پژوهش آنان، دو نکته قابل تأمل به چشم می‌خورد. نخست، در این پژوهش به وضوح مشخص نشده است که منظور از حالت «جدانویسی»، نگارش با فاصله کامل است یا حالت نیم فاصله نویسی؛ یا اینکه این دو حالت نگارشی یکی در نظر گرفته شده است. اگر مبنای پژوهشگران دستور خط مصوب فرهنگستان باشد، در این حالت منظور از جدانویسی حالت نیم فاصله نویسی یا نزدیک نویسی است. نکته دوم این است که آیا در عنوان‌های انتخاب شده به عنوان حجم نمونه، واژگان شامل حالت «نیم فاصله نویسی» هم مشاهده شده است یا خیر و اگر وجود داشته، آیا در زمان جستجو در پایگاه اطلاعات علمی فارسی این حالت نگارشی هم جستجو و بررسی شده است یا خیر.

در نهایت «هماوندی، نوروزی و حسینی بهشتی» (۱۳۹۷) در پژوهشی به روش پیمایشی-تحلیلی و با استفاده از شیوه مشاهده مستقیم، به تشریح مشکلات عمده نوشتاری و معنایی زبان فارسی در استفاده از محیط‌های اطلاعاتی و تعیین میزان انطباق و توجه به این ویژگی‌ها هنگام جستجو و بازیابی در پایگاه‌های اطلاعاتی فارسی پرداختند. یافته‌های آنان نشان داد پایگاه‌های اطلاعاتی فارسی نسبت به ویژگی‌های نوشتاری و معنایی زبان فارسی، همچون پیوسته نویسی و جدانویسی، گوناگونی جمع‌ها، واژگان دخیل و معادل آنها توجه کافی ندارند و بسیاری از ویژگی‌های آن را در مراحل ذخیره‌سازی و پردازش اطلاعات نادیده گرفته‌اند. در ارتباط با پژوهش «هماوندی، نوروزی و حسینی بهشتی» موارد قابل تأملی به چشم می‌خورد که پیش از این درباره پژوهش «جوکار و عبدالمهی» (۱۳۸۸)، «گل تاجی و بذرگر» (۱۳۸۹)، و «آخشیک و فتاحی» (۱۳۹۱) به آن پرداخته شد. مواردی چون انتخاب و اکتفا به چند مصداق برای چالش‌های مورد بررسی، آن هم بدون ذکر دلایل انتخاب آنها.

مرور پژوهش‌های پیشین نشان داد پژوهشگران برای تعیین میزان توجه سامانه‌های بازیابی اطلاعات به چالش‌های خط فارسی، از جمله مسئله پیوسته‌نویسی و جدانویسی، گاه با ذکر تعداد محدودی مصداق (یک یا دو مصداق) آن هم از قلمرو موضوعی محدود، وضعیت سامانه‌های بازیابی اطلاعات را بررسی کرده‌اند؛ درحالی‌که با توجه به پوشش قلمروهای موضوعی متعدد در سامانه‌های بازیابی اطلاعات مورد مطالعه آنها، با در نظر گرفتن این جنبه می‌شد با جامعیت بیشتر و در نتیجه به شکل دقیق‌تر درباره میزان تأثیر چالش‌های نگارشی قضاوت و راهکارهایی را برای اصلاح یا بهبود آنها ارائه کرد. علاوه بر خلاً فوق، پژوهشگران پس از بازیابی رکوردها در سامانه بازیابی اطلاعات، یکسانی فراوانی تعداد مدارک بازیابی شده در ریخت‌های مختلف یک واژه را نشان از رفع آن چالش در سامانه‌های بازیابی اطلاعات قلمداد کرده‌اند. در توضیح این شیوه پژوهشگران نیز باید گفت کنترل شکل‌های نگارشی یک واژه در بخش‌های یک رکورد (عنوان، چکیده و واژگان کلیدی) امری ضروری است و حتی احتمال دارد فقط یک حالت نگارشی در مدرک رخ داده باشد و سامانه‌های بازیابی به‌گونه‌ای طراحی شده‌اند که قادر به بهنجارسازی شکل‌های نگارشی یک واژه باشند.

در بخش ترکیبات دستور خط فارسی مصوب فرهنگستان، ۲۳ قاعده درباره موقعیت‌های رخداد جدانویسی و پیوسته‌نویسی ذکر شده است؛ در صورتی که مرور پژوهش‌های پیشین نشان می‌دهد پژوهشگران بدون ذکر چگونگی و دلایل انتخاب واژگان بیان‌کننده چالش پیوسته‌نویسی و جدانویسی، بعضاً با در نظر گرفتن یک یا دو مصداق‌ها (کلیدواژه‌ها) به بررسی این چالش و همچنین میزان تأثیر آن بر سامانه‌های بازیابی اطلاعات پرداخته‌اند. در صورتی که به نظر می‌رسد انتخاب مصداق‌ها برای هر قاعده می‌تواند تأثیرگذاری متفاوتی بر سامانه‌های بازیابی اطلاعات داشته باشد. تبیین این مسائل به منزله شکاف‌های پژوهشی این قلمروست که سنجش آن می‌تواند سودمند باشد. پس از مرور تحلیلی - انتقادی پیشینه‌های این مقاله، نویسندگان بر این باورند که

پژوهش مشابهی با پژوهش حاضر مشاهده نشد و ضرورت انجام این پژوهش، بیش از پیش احساس می‌شود.

پیشینه‌های خارجی

در خارج ایران نیز پیرامون سایر زبان‌ها و مشکلات آنها در مواجهه با محیط‌های الکترونیکی مطالعات مشابه گوناگونی انجام شده است که در ادامه به برخی از آنها اشاره می‌شود.

«ژانگ و لین»^۱ (۲۰۰۷) در پژوهشی پیمایشی - مقایسه‌ای ویژگی‌های پشتیبانی چندزبانه به وسیله موتورهای جستجو شبکه اینترنت را بررسی کردند. یافته‌های آنان نشان داد موتورهای جستجوی گوگل، EZ2Find و Onlinelink در بین بسیاری از موتورهای جستجو مجهز به ویژگی‌های پشتیبانی چندزبانه، وضعیت مطلوب‌تری دارند.

«لازارینیس»^۲ (۲۰۰۷) در پژوهشی با رویکرد پیمایشی، قابلیت‌های جستجوی وب‌سایت‌های الکترونیکی تجاری را درباره زبان‌های غیرانگلیسی (مطالعه موردی یونانی) بررسی کرد. نتایج او نشان داد موتورهای جستجوی محلی به ریخت‌شناسی سؤال‌ها (کلیدواژه‌ها) توجهی نشان نمی‌دهد که نهایت این امر به شکست جستجوی کاربر می‌انجامد.

«لواندوفسکی»^۳ (۲۰۰۸) در پژوهش خود با رویکرد پیمایشی توانایی موتورهای جستجوی پراستفاده و اصلی از جمله گوگل، یاهو، ام‌اس‌ان^۴ و اسک^۵ در تشخیص و تمایز میان مدارک به زبان آلمانی از پیشینه‌هایی با زبان انگلیسی را بررسی کرد. نتایج پژوهش او نشان داد در موتورهای کاوش گوگل و ام‌اس‌ان، وقتی نتایج به زبان خاصی

1. Zhang & Lin
2. Lazarinis
3. Lewandowski
4. MSN
5. Ask

محدود می‌شود، کاربر با مشکلاتی روبه‌رو می‌شود، درحالی‌که هیچ‌یک از موتورهای کاوش در بازیابی نتایج به زبان صفحه رابط کاربر (زبان انتخابی) با مشکل مواجه نمی‌شوند. افزون بر این، نتایج او نشان داد استفاده از صفحه میانی به زبان بومی در جستجو و بازیابی اطلاعات در برخی مواقع اثرگذاری بهتری دارد و همچنین راهبرد محدودیت زبانی، همیشه در بهبود جستجو تأثیرگذار نیست.

در پژوهشی دیگر «همو»^۱ با رویکرد پیمایشی یک قالب کاری برای افزایش کارایی موتورهای کاوش برای متون عربی دار و همچنین فاقد اعراب‌گذاری از طریق روش‌های گسترش سؤال (کلیدواژه) ارائه کرد. نتایج پژوهش او نشان داد گسترش سؤال (کلیدواژه) بر بهبود جستجو و بازیابی متون عربی تأثیرگذار است و کارایی موتورهای کاوش با استفاده از ابزارهای پیشرفته پردازش زبان طبیعی، افزایش می‌یابد.

مرور پژوهش‌های خارجی نیز نشان می‌دهد این پژوهش‌ها اغلب به بررسی توانمندی‌ها و ضعف‌های سامانه‌های جستجو، با هدف شناخت و ارائه راهکارهایی برای اصلاح چالش‌های زبانی پرداخته‌اند. یافته‌های حاصل از آنها نشان می‌دهد ریخت‌شناسی واژه‌ها و عبارت‌های جستجو بر بازیابی مدارک مؤثر است و ابزارهای جستجو در هنگام بازیابی نتایج بر شکل کلیدواژه‌های جستجو شده تکیه می‌کنند که این امر در نهایت می‌تواند به شکست جستجوی کاربر منتهی شود.

روش‌شناسی پژوهش

پژوهش حاضر از نظر هدف در ردیف پژوهش‌های کاربردی و از جنبه طرح پژوهش از نوع تحلیل محتواست. جامعه آماری شامل مجموعه مدارک (مقاله‌های علمی - پژوهشی و علمی - ترویجی) ذخیره شده با امکان دسترسی به نسخه تمام‌متن آنها در

پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور»^۱، پایگاه «مرکز اطلاعات علمی جهاد دانشگاهی»^۲ و پایگاه «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»^۳ است. شایان ذکر است، این سه پایگاه اطلاعات علمی فارسی به دلیل پوشش نسبتاً جامع متون علمی قلمروهای موضوعی مختلف - که در این پژوهش دسته‌بندی قلمروهای موضوعی «سامانه ارزیابی نشریات علمی»^۴ مورد استفاده قرار گرفته است و شامل قلمروهای دام‌پزشکی، علوم انسانی، فنی و مهندسی، علوم پایه، منابع طبیعی و کشاورزی، هنر و معماری است - و همچنین، به دلیل دارا بودن بیشترین فراوانی در بین پایگاه‌های اطلاعات علمی مطالعه شده در پژوهش‌های قلمرو بازایی اطلاعات، به عنوان مهم‌ترین پایگاه‌های اطلاعات علمی فارسی تلقی و محمل انتخاب مدارک شامل مصداق‌های «کلیدواژه‌ها» (واژگان) با قابلیت پیوسته‌نویسی، نزدیک‌نویسی و جدانویسی قرار گرفتند. براین اساس، تعداد کل مدارک بازایی شده از سه پایگاه اطلاعات علمی فارسی مورد مطالعه ۲۹۱۴۲ مدرک است که به عنوان جامعه آماری انتخاب شد.

حجم نمونه نیز با توجه به وجود سه پایگاه اطلاعات علمی فارسی مورد مطالعه و نیاز به مقایسه شش قلمرو موضوعی و همچنین ده قاعده انتخابی (از مجموع قاعده‌های گزینش شده بخش «ترکیبات» فرهنگستان) با استفاده از روش نمونه‌گیری طبقه‌ای نسبتی و فرمول کوکران برای جامعه معین، با خطای ۰/۰۵ و از طریق فرمول
$$n = \frac{NZ^2pq}{Nd^2 + Z^2pq}$$
 (دلور، ۱۳۹۰) تعداد ۳۸۰ مدرک تعیین شد؛ اما با توجه به هدف پژوهش، لازم بود تا این تعداد مدارک به عنوان نمونه در سه سطح، یعنی سه پایگاه اطلاعات علمی فارسی مورد مطالعه، شش قلمرو موضوعی و در هر قلمرو موضوعی نیز بین قاعده‌های دهگانه آنها توزیع شود. بنابراین، پس از توزیع حجم نمونه در سه سطح

1. Magiran; (www.magiran.com)

2. Scientific information Database (SID); (www.sid.ir)

3. Regional information center for science and technology (RICeST); (www.ricest.ac.ir)

۴. قابل دسترس در: <http://journals.msrt.ir>

گفته شده، تعداد مدارک در سطح سوم یعنی به ازای هر قاعده، ۱/۷۵ مدرک به دست می‌آید که این مقدار به عدد ۲ گرد شد. با این عمل، حجم نمونه نهایی به ۴۳۲ مدرک افزایش یافت. بدین شکل، با توزیع نمونه در سه سطح گفته شده، به ازای هر پایگاه اطلاعات علمی فارسی مورد مطالعه ۱۴۴ مدرک، به ازای هر قلمرو موضوعی ۲۴ مدرک و به ازای هر قاعده در هر قلمرو موضوعی نیز ۲ مدرک انتخاب شد.

پس از تعیین حجم نمونه مدارک، با مینا قراردادن بخش ترکیبات دستور خط فارسی مصوب فرهنگستان و با همکاری و نظر استادان رشته‌های علم اطلاعات و دانش‌شناسی و زبان‌شناسی، ۲۳ قاعده بخش ترکیبات ارزیابی شد. برخی از قاعده‌ها به دلایلی همچون شمی بودن و دشواری تشخیص آنها (قاعده‌های ۳، ۵، ۶ و ۷ از مجموعه موارد پیوسته نویسی الزامی و قاعده‌های ۶، ۱۴ و ۱۵ مجموعه موارد جدانویسی الزامی) و همچنین عدم رویداد مصداق‌ها (کلیدواژه‌ها) متناسب با آنها در اصطلاح‌نامه‌ها و فرهنگ‌های تخصصی مورد استفاده یا عدم بازیابی مدرک در زمان جستجوی آنها در پایگاه‌های اطلاعات علمی فارسی مورد مطالعه (قاعده‌های ۴، ۵ و ۸ از موارد الزامی پیوسته نویسی و قاعده‌های ۵، ۱۳ و ۷ از موارد جدانویسی الزامی) از مجموع قاعده‌های کنار گذاشته شد. در نهایت، ده قاعده از ۲۳ قاعده دستور خط فارسی مصوب فرهنگستان برای ساخت سیاهه و آرسی شامل مصداق‌ها (کلیدواژه‌ها) دارای شکل نگارشی پیوسته نویسی و جدانویسی انتخاب شد که از این پس، قاعده‌های دهگانه خوانده می‌شود. در بین ده قاعده انتخابی، مواردی که انتخاب شکل نگارشی به نویسنده واگذار شده بود، بر مبنای کتاب «فرهنگ املائی خط فارسی»^۱ چاپ ۱۳۹۵ عمل شد. لازم به ذکر است، از بین قاعده‌های دهگانه، دو قاعده ۱ و ۵ به منظور عملیاتی‌سازی سنجش آنها، به دو قاعده فرعی مجزا تفکیک و با حروف «الف» و «ب» از

۱. صادقی، علی‌اشرف؛ وزندی مقدم، زهرا. (۱۳۹۵). فرهنگ املائی خط فارسی براساس دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی. تهران: فرهنگستان زبان و ادب فارسی. قابل دسترس در وبگاه فرهنگستان زبان و ادب فارسی:

هم متمایز شدند (جدول ۱).

جدول ۱. قاعده‌های دهگانه دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی

قاعده ۱ (الف)	هنگامی که ترکیب پردندانه شود، ترکیب جدا [نزدیک‌نویسی] نوشته می‌شود، مثل آب‌شستگی
قاعده ۱ (ب)	هنگامی که ترکیب طولانی شود، ترکیب جدا [نزدیک‌نویسی] نوشته می‌شود، مثل سوراخ‌کاری
قاعده ۲	رسیدن حروف مشابه یا یکسان و هم‌مخرج به هم موجب جدانویسی [نزدیک‌نویسی] می‌شود، مثل آب‌بند، سیم‌پیچ
قاعده ۳	وقتی جزء دوم با الف آغاز شود، موجب جدانویسی [نزدیک‌نویسی] می‌شود، مانند کم‌احساس، هم‌اسم
قاعده ۴	وقتی جزء دوم با «آ» آغاز شود و تک‌هجایی باشد، موجب جدانویسی [نزدیک‌نویسی] می‌شود و در صورتی که جزء دوم بیش از یک هجا داشته باشد، از قاعده خاصی پیروی نمی‌کند و گاهی جدا و گاهی پیوسته نوشته می‌شود که در این حالت شکل ارجح واژه بر اساس فرهنگ املائی خط فارسی انتخاب می‌شود، مثل بتن‌آرمه، زبان‌آموزی
قاعده ۵ (الف)	ختم‌شدن واژه اول در ترکیب به حروف پیوند ناپذیر موجب جدانویسی [نزدیک‌نویسی] می‌شود، مثل کشتارگاه.
قاعده ۵ (ب)	ختم‌شدن واژه اول در ترکیب به «های» غیرملفوظ موجب جدانویسی [نزدیک‌نویسی] می‌شود، مثل قفسه‌سینه.
قاعده ۶	ترکیب‌های اضافی (موصوف و صفت/ مضاف و مضاف‌الیه) جدا نوشته می‌شوند، مثل جسم‌زرد
قاعده ۷	وقتی یک جزء واژه مرکب عدد باشد، موجب جدانویسی [نزدیک‌نویسی] ترکیب می‌شود، مثل هشت‌بهشت
قاعده ۸	واژگان مرکبی که از ترکیب با پیشوند ساخته می‌شود، همیشه جدا [نزدیک‌نویسی] نوشته می‌شود، مگر مرکب‌هایی که با پیشوندهای «به»، «بی» و «هم» با رعایت استثناهایی - صفحات ۲۲-۲۳ دستور خط فارسی - ساخته می‌شود. در این حالت نیز در صورت مواجهه با واژگانی که اختیار به نویسنده واگذار شده باشد، از فرهنگ املائی خط فارسی استفاده می‌شود، مثل همبندی، بی‌حسی موضعی
قاعده ۹	واژگان مرکبی که از ترکیب با پسوند ساخته می‌شود، همیشه پیوسته نوشته می‌شود، مگر هنگامی که مطابق با قاعده ۲، ۴ و ۷ ساخته شوند، مثل نوسانگر، شالیزار
قاعده ۱۰	یک جزء واژه مرکب صفت فاعلی یا مفعولی باشد، جدا [نزدیک‌نویسی] نوشته می‌شود، مثل اجل‌رسیده، تنظیم‌کننده

در مرحله بعد لازم بود به تفکیک قاعده‌های دهگانه، مصداق‌ها (واژگان یا کلیدواژه‌های انتخابی) مورد نیاز هر قاعده انتخاب شود. بدین منظور، از اصطلاح‌نامه‌های تخصصی و در صورت نبود اصطلاح‌نامه در هر یک از قلمروهای موضوعی (تنها در دام پزشکی) از فرهنگ تخصصی آن قلمرو استفاده شد. پس از یافتن هر یک از اصطلاحات یا واژگان مورد نظر (به‌عنوان مصداق مورد نظر)، اصطلاح یا واژه در هر سه پایگاه اطلاعات علمی فارسی جستجو شد و در صورت بازیابی بیش از یک مدرک به ازای هر مصداق و همچنین رویداد آن در هر سه پایگاه اطلاعات علمی فارسی، اصطلاح یا لغت برای قاعده مورد نظر انتخاب شد. در نهایت، مصداق‌های انتخابی برای هر قاعده از قاعده‌های دهگانه به تفکیک در یک سیاهه واری قرار گرفت. لازم به ذکر است، در این مرحله به دلیل حجیم بودن اصطلاح‌نامه‌ها و فرهنگ تخصصی، عملاً مرور تمامی صفحات آنها امکان‌پذیر نبود. بدین منظور، ابتدا اصطلاح‌نامه‌ها و فرهنگ تخصصی بر حسب حروف الفبا و مقدار حجم لغات هر حرف الفبا، به سه سطح بزرگ، متوسط و کوتاه تقسیم شد. سپس، از هر یک از سطوح تعیین شده به روش تصادفی ساده، یک حرف الفبا انتخاب و آنگاه برای هر قلمرو موضوعی، فرایند جستجو برای مصداق‌های مورد نظر بر حسب قاعده‌های دهگانه انجام شد.

در گام بعد، با مدنظر قرار دادن سیاهه واری تهیه شده در مرحله قبل، به هر یک از پایگاه‌های اطلاعاتی مورد مطالعه مراجعه شد و مصداق‌ها با تمام شکل‌های نگارشی در کادر جستجوی ساده سه پایگاه اطلاعات علمی فارسی مورد مطالعه جستجو گردید.^۱ شایان ذکر است، به دلیل عدم پشتیبانی و شناسایی پایگاه‌های اطلاعاتی مورد مطالعه (به جز پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری») از حالت نیم‌فاصله در هنگام جستجو و در نتیجه عدم بازیابی مدارک دارای شکل نگارشی نزدیک‌نویسی، این حالت تنها در این بخش از پژوهش از فهرست مجموع حالت‌های

۱. به منظور اجتناب از تغییراتی که ممکن است به علت روزآمدسازی پایگاه‌های اطلاعاتی پیش‌آید، فرایند جستجویی یک روز انجام شد.

مورد مطالعه کنار گذاشته شد.

در مرحله پیشین، از بین مجموع مدارک بازیابی شده، به ازای هریک از حالت‌های نگارشی (پیوسته‌نویسی یا جدانویسی)، تعداد یک مدرک و در صورتی که هیچ مدرکی با حالت نگارشی مورد نظر در پایگاه اطلاعات علمی فارسی بازیابی نمی‌شد، هر دو مدرک از حالت نگارشی دیگر به عنوان نمونه انتخاب می‌شد (در مجموع دو مدرک برای هر قاعده از هر قلمرو موضوعی). چنان‌که می‌دانید، در روش نمونه‌گیری تصادفی ساده، احتمال و شانس انتخاب شدن برای همه اعضای جامعه برابر و مستقل از یکدیگر است. این نوع نمونه‌گیری را به سه شیوه می‌توان انجام داد: الف) قرعه‌کشی ب) جدول اعداد تصادفی ج) استفاده از نرم‌افزارهای آماری (ولیکسانی و سرافراز، ۱۳۹۴). از این رو، پژوهشگر برای انتخاب دو مدرک از بین مجموع مدارک بازیابی شده برای هر مصداق، از روش نمونه‌گیری تصادفی ساده و از نرم‌افزار تولید اعداد تصادفی^۱ استفاده کرد. شایان ذکر است، بدین شکل فرایند انتخاب و گردآوری ۴۳۲ مدرک مورد نیاز به عنوان حجم نمونه صورت پذیرفت.

در گام آخر، به منظور سنجش و ثبت قابلیت پایگاه‌های اطلاعات علمی فارسی از نظر بازیابی مدارک با هریک از شکل‌های نگارشی مصداق‌ها اعم از پیوسته‌نویسی یا جدانویسی، ابتدا به هریک از ۴۳۲ مدرک انتخاب شده به عنوان نمونه در پایگاه‌های اطلاعات علمی فارسی مراجعه شد و فرایند تحلیل محتوا در عنوان، چکیده و واژگان کلیدی هر مدرک انجام و پس از اطمینان از عدم رخداد همزمان چند شکل نگارشی مصداق در بخش چکیده، عنوان و واژگان کلیدی پایگاه اطلاعات علمی فارسی مورد نظر، مدارکی که با هر دو حالت نگارشی مصداق‌ها اعم از پیوسته‌نویسی (سره‌نویسی) و جدانویسی (فاصله کامل) قابل بازیابی بود، این ویژگی به عنوان حالت «گسترش‌پذیری جستجو» برای پایگاه اطلاعات علمی فارسی مورد نظر در نظر گرفته می‌شد؛ در غیر این صورت، به عنوان حالت «گسترش‌ناپذیری جستجو» ثبت گردید.

۱. این نرم‌افزار از طریق آدرس <https://www.spss-iran.com> قابل دسترسی است.

یافته‌ها

چنان‌که جدول ۲ نشان می‌دهد، در مجموع پایگاه‌های مورد بررسی تنها ۲۳ مدرک (۳/۵٪) با حالت گسترش‌پذیری جستجو‌بازیابی شدند. در مقابل، ۴۰۹ مدرک (۷/۹۴٪) با حالت گسترش‌ناپذیری جستجو‌بازیابی شدند.

جدول ۲. قابلیت پایگاه‌های اطلاعات علمی فارسی از نظر گسترش جستجو و بازیابی مدارک به صورت پیوسته نویسی و جدانویسی

X ²	(df=۲)	گسترش جستجو			پایگاه	SID
		کل	گسترش ناپذیری جستجو	گسترش پذیری جستجو		
		۱۴۴	۱۴۴	۰	فراوانی مشاهده شده	
		۱۴۴	۱۳۶/۳	۷/۷	فراوانی مورد انتظار	
		۱۰۰	۱۰۰	۰	درصد در درون پایگاه اطلاعات علمی فارسی	
		۱۴۴	۱۳۱	۱۳	فراوانی مشاهده شده	MAGIRAN
		۱۴۴	۱۳۶/۳	۷/۷	فراوانی مورد انتظار	
۰/۰۰۲	۱۲/۷۶۷**	۰/۰۰۲	۱۰۰	۹	درصد در درون پایگاه اطلاعات علمی فارسی	
		۱۴۴	۱۳۴	۱۰	فراوانی مشاهده شده	RICEST
		۱۴۴	۱۳۶/۳	۷/۷	فراوانی مورد انتظار	
		۱۰۰	۹۳/۱	۶/۹	درصد در درون پایگاه اطلاعات علمی فارسی	
		۴۳۲	۴۰۹	۲۳	فراوانی مشاهده شده	کل
		۴۳۲	۴۰۹	۲۳	فراوانی مورد انتظار	
		۱۰۰	۹۴/۷	۵/۳	درصد در درون پایگاه‌های اطلاعاتی	

*P<۰/۰۵ **P<۰/۰۱

چنان‌که می‌دانید، آزمون مجذور خی را می‌توان هم برای رابطه و هم تفاوت بین متغیرهای اسمی یا مقوله‌هایی به کار برد که دو یا بیش از دو ارزش داشته باشند. این نوع آزمون تنها می‌تواند در رابطه با تعداد یا فراوانی‌ها به کار رود (دلاور، ۱۳۹۰). از این رو، برای بررسی معناداری تفاوت میان فراوانی در حالت گسترش‌پذیری جستجو و گسترش‌ناپذیری جستجوی مصداق‌ها مدارک از آزمون مجذور خی استفاده شد. با توجه به مقدار مجذور خی به دست آمده ($\chi^2_{(12/767)} = 0.002$ ، $P = 0$)، تفاوت معناداری میان پایگاه‌های اطلاعات علمی فارسی در دو دسته گسترش‌پذیری جستجو و گسترش‌ناپذیری جستجوی مصداق‌های مدارک وجود داشت. پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی» بیشترین فراوانی را در گسترش‌ناپذیری جستجوی مصداق‌های مدارک و همچنین کمترین فراوانی را در گسترش‌پذیری جستجوی مصداق‌های مدارک به خود اختصاص داده بود؛ و در مقابل، پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» بیشترین فراوانی را در گسترش‌پذیری جستجوی مصداق‌های مدارک و کمترین فراوانی را در گسترش‌ناپذیری جستجوی مصداق‌های مدارک داشت.

علاوه بر یافته بالا، نتایج به دست آمده از توزیع فراوانی قابلیت پایگاه‌های اطلاعات علمی فارسی از نظر گسترش جستجو و بازیابی مدارک به صورت پیوسته‌نویسی و جدانویسی بر حسب قاعده‌های دهگانه فرهنگستان زبان و ادب فارسی نشان داد، پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» در مصداق‌های مربوط به قاعده‌های ۱، ۲، ۵ و ۹ قادر است با هریک از شکل‌های نگارشی مصداق‌ها اعم از پیوسته‌نویسی و جدانویسی مدارک مورد نظر را دوباره بازیابی کند؛ بدین شکل که پایگاه اطلاعات علمی فارسی مذکور از کل ۱۳ مدرک به دست آمده با حالت گسترش‌پذیری جستجو، در ۹ مدرک شامل شکل نگارشی جدانویسی مصداق، قادر به بازیابی مجدد این مدارک حتی با حالت جستجوی پیوسته‌نویسی مصداق است (گسترش پیوسته‌نویسی به جدانویسی) و در ۴ مدرک دیگر که دارای شکل نگارشی پیوسته‌نویسی

از مصداق‌ها بود، پایگاه اطلاعات علمی فارسی قادر به بازیابی دوباره مورد نظر حتی با حالت جستجوی جدانویسی مصداق‌ها بود (گسترش جدانویسی به پیوسته‌نویسی). در مقابل، پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» در قاعده‌های بیشتری (۱، ۲، ۳، ۸، ۹ و ۱۰)، از نظر توانایی بازیابی مدارک با هریک از شکل‌های نگارشی در دامنه‌ای محدودتر نسبت به پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» جای داشت؛ بدین ترتیب که در کل مدارک به دست آمده با حالت گسترش‌پذیری جستجو (۱۰ مدرک) با بررسی عنوان، چکیده و واژگان کلیدی در هر ده مدرک، فقط شکل نزدیک‌نویسی مصداق‌ها روی داده بود و پایگاه اطلاعات علمی فارسی مذکور در صورت جستجوی مصداق‌ها با حالت جدانویسی، علاوه بر بازیابی مدارک شامل شکل نگارشی جدانویسی، مدارک دارای شکل نگارشی نزدیک‌نویسی را هم بازیابی می‌کرد. شایان ذکر است، پایگاه اطلاعات علمی فارسی «مرکز اطلاعات علمی جهاد دانشگاهی» به دلیل فراوانی صفر برای گسترش‌پذیری جستجو و در نتیجه ناتوانی در بازیابی مدارک با هریک از شکل‌های نگارشی مصداق‌ها، در این قسمت گزارش نشده است (جدول ۳).

جدول ۳. قابلیت پایگاه‌های اطلاعات علمی در گسترش‌پذیری جستجو به تفکیک قاعده‌های دهگانه

پایگاه‌های اطلاعاتی	شماره قاعده	قابلیت پایگاه	فراوانی درصد	
MAGIRAN	الف ۱	گسترش پیوسته به جدا	۲ / ۱۵/۴	
	ب ۱	گسترش پیوسته به جدا	۲ / ۱۵/۴	
	۲	گسترش پیوسته به جدا	۳ / ۲۳	
	الف ۵	گسترش جدا به پیوسته	۲ / ۱۵/۴	
	۹		گسترش پیوسته به جدا	۲ / ۱۵/۴
			گسترش جدا به پیوسته	۲ / ۱۵/۴
	کل		۱۳ / ۱۰۰	

پایگاه‌های اطلاعاتی	شماره قاعده	قابلیت پایگاه	فراوانی	درصد
RICEST	الف ۱	گسترش جدا به نیم فاصله	۲	۲۰
	۲	گسترش جدا به نیم فاصله	۱	۱۰
	۳	گسترش جدا به نیم فاصله	۲	۲۰
	۸	گسترش جدا به نیم فاصله	۲	۲۰
	۹	گسترش جدا به نیم فاصله	۲	۲۰
	۱۰	گسترش جدا به نیم فاصله	۱	۱۰
کل			۱۰	۱۰۰

علاوه بر یافته بالا، مقایسه شکل نگارشی مصداق‌ها در بخش عنوان، چکیده و واژگان کلیدی نسخه پی.دی.اف^۱ هر مدرک با اطلاعات ارائه شده از همان مدرک در پایگاه‌های اطلاعات علمی فارسی نشان داد، شکل نگارشی بخش قابل توجهی از مصداق‌های مدارک تغییر کرده است (جدول ۴).

جدول ۴. توزیع فراوانی تغییرات شکل نگارشی مصداق‌ها (کلیدواژه‌های انتخابی) از مدرک به پایگاه‌های اطلاعات علمی فارسی

پایگاه	تغییرات شکل نگارشی مصداق (از مدرک به پایگاه اطلاعاتی)	فراوانی	درصد
SID	پیوسته به جدا	۱	۰/۷
	نیم فاصله به پیوسته	۱	۰/۷
	پیوسته و جدا به جدا	۱	۰/۷
	پیوسته، نیم فاصله و جدا به پیوسته و جدا	۱	۰/۷
	نیم فاصله به پیوسته، نیم فاصله و جدا	۲	۱/۴
	نیم فاصله به نیم فاصله و جدا	۲	۱/۴

درصد	فراوانی	تغییرات شکل نگارشی مصداق (از مدرک به پایگاه اطلاعاتی)	
		پایگاه	
۲/۸	۴	پیوسته و جدا به پیوسته و جدا	
۴/۹	۷	نیم فاصله و جدا به جدا	
۲۲/۹	۳۳	پیوسته به پیوسته	
۲۸/۵	۴۱	نیم فاصله به جدا	
۳۵/۴	۵۱	جدا به جدا	
۱۰۰	۱۴۴	کل	
۰/۷	۱	نیم فاصله و جدا به پیوسته و جدا	MAGIRAN
۱/۴	۲	پیوسته به جدا	
۱/۴	۲	نیم فاصله و جدا به جدا	
۲/۸	۴	پیوسته به پیوسته و جدا	
۴/۹	۷	پیوسته و جدا به پیوسته و جدا	
۵/۶	۸	نیم فاصله به پیوسته	
۵/۶	۸	نیم فاصله به پیوسته و جدا	
۲۳/۶	۳۴	پیوسته به پیوسته	
۲۳/۶	۳۴	نیم فاصله به جدا	
۳۰/۶	۴۴	جدا به جدا	
۱۰۰	۱۴۴	کل	
۰/۷	۱	نیم فاصله به نیم فاصله و پیوسته	RICEST
۰/۷	۱	پیوسته و نیم فاصله به جدا	
۰/۷	۱	نیم فاصله به نیم فاصله و جدا	
۲/۸	۴	پیوسته به جدا	
۲/۸	۴	نیم فاصله و جدا به جدا	
۲/۸	۴	نیم فاصله به پیوسته، نیم فاصله و جدا	
۳/۵	۵	پیوسته و جدا به پیوسته و جدا	
۴/۲	۶	نیم فاصله به پیوسته و جدا	

درصد	فراوانی	تغییرات شکل نگارشی مصداق (از مدرک به پایگاه اطلاعاتی)	
		پایگاه	
۶/۳	۹	نیم فاصله به پیوسته	
۶/۳	۹	نیم فاصله به نیم فاصله	
۱۶/۷	۲۴	نیم فاصله به جدا	
۲۴/۳	۳۵	پیوسته به پیوسته	
۲۸/۵	۴۱	جدا به جدا	
۱۰۰	۱۴۴	کل	

با توجه به جدول ۴ در پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی» از مجموع ۱۴۴ مدرک، در ۸۸ مدرک شکل نگارشی مصداق‌ها تغییری نکرده بود؛ اما در ۵۶ مدرک دیگر، شکل نگارشی مصداق‌ها تغییر داشت. این وضعیت تقریباً با همین نسبت در پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» نیز مشاهده می‌شود؛ بدین ترتیب که در ۸۵ مدرک شکل نگارشی مصداق‌ها بدون تغییر، اما در ۵۹ مدرک دیگر شکل نگارشی مصداق‌ها تغییر یافته بود. پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» متفاوت با دو پایگاه اطلاعات علمی فارسی دیگر بود؛ به گونه‌ای که در ۹۰ مدرک، تغییری در شکل نگارشی مصداق‌ها روی نداده بود و تنها در ۵۴ مدرک شکل نگارشی مصداق‌ها بین نسخه پی.دی.اف مدرک با پایگاه اطلاعات علمی فارسی متفاوت بود.

بحث و نتیجه‌گیری

به طور خلاصه، یافته‌های پژوهش حاضر نشان داد پایگاه‌های اطلاعات علمی فارسی در هنگام جستجوی شکل نگارشی نزدیک نویسی و در نتیجه بازیابی مدارک دارای صرفاً شکل نگارشی نزدیک نویسی پشتیبانی نمی‌کنند؛ بدین ترتیب که هنگام جستجو با این حالت نگارشی در پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» همان

مدارک دارای شکل نگارشی جدانویسی بازیابی می‌شوند و مدارک دارای صرفاً شکل نگارشی نزدیک نویسی بازیابی نمی‌شدند. در پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی»، علاوه بر بازیابی مدارک دارای شکل نگارشی جدانویسی، مدارک دارای یکی از اجزای مصداق‌ها (کلیدواژه‌ها) هم بازیابی می‌شدند و این امر موجب ریزش کاذب^۱ قابل توجه در مدارک بازیابی شده می‌گردید. این یافته با یافته‌های پژوهش «گل‌تاجی و بذرگر» (۱۳۸۹) همسوییست. در این پژوهش، پژوهشگران یکسان بودن تعداد مدارک بازیابی شده با حالت‌های جدانویسی و نزدیک نویسی را به بهنجارسازی شکل نگارشی از سوی پایگاه‌های اطلاعاتی نسبت داده‌اند؛ درحالی‌که یافته‌های پژوهش حاضر نشان می‌دهد از دلایل احتمالی یکسان بودن تعداد کل مدارک بازیابی شده، می‌تواند به عدم پشتیبانی پایگاه اطلاعات علمی فارسی از شکل نگارشی نزدیک نویسی و یا به رخداد همزمان چند شکل نگارشی مصداق در بخش‌های مختلف یک مدرک اعم از عنوان، چکیده و واژگان کلیدی بازگردد.

علاوه بر این، نتایج نشان داد میان قابلیت پایگاه‌های اطلاعات علمی فارسی از نظر بازیابی مدارک با هر یک از شکل‌های نگارشی اعم از پیوسته نویسی و جدانویسی تفاوت وجود دارد. بدین ترتیب که پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» با بیشترین فراوانی در گسترش‌پذیری جستجو نسبت به سایر پایگاه‌های اطلاعات علمی فارسی از عملکرد مطلوب‌تری برخوردار است؛ به‌گونه‌ای که در مدارک به دست آمده با حالت گسترش‌پذیری جستجو در این پایگاه اطلاعات علمی فارسی، با جستجوی مصداق‌های آنها با هر یک از حالت‌های جدانویسی یا پیوسته نویسی، مدارک مورد نظر دوباره قابل بازیابی است. در مقابل، پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی» با کمترین فراوانی در گسترش‌پذیری جستجوی مصداق‌های

۱. خروجی و برونداد غیرمرتبطی که در نتیجه اجرای راهبرد کاوش در سیستم بازیابی اطلاعات تولید می‌شود. محاسبه ریزش کاذب از طریق فرمول زیر (یوسفی، ۱۳۷۶): مدارک بازیابی شده مرتبط - تعداد کل مدارک بازیابی شده = ریزش کاذب.

مدارک نسبت به دو پایگاه اطلاعات علمی فارسی دیگر، از وضعیت نامطلوبی برخوردار بود؛ بدین ترتیب که با اعمال هریک از تغییرها در شکل نگارشی به هنگام جستجو، مدارک مورد نظر بازاریابی نمی‌شد و تنها جستجوی مصداق‌ها با همان حالت ثبت شده در مدارک به بازاریابی دوباره آنها می‌انجامید.

از این رو، نتایج کلی به دست آمده این بخش از پژوهش با یافته‌های پژوهش‌های «مرتضایی» (۱۳۸۱)، «عبدالهی و جوکار» (۱۳۸۸)، «گل تاجی و بذرگر» (۱۳۸۹)، «آخشیک و فتاحی» (۱۳۹۱)، «هماوندی و دیگران» (۱۳۹۷) هم‌راستاست؛ بدین شکل که نتایج این بخش از پژوهش مهر تأییدی است بر یافته‌های «مرتضایی» (۱۳۸۱) مبنی بر اینکه استاندارد نبودن شکل نوشتاری واژگان در عدم مطلوبیت و جامعیت نتایج جستجو تأثیر می‌گذارد. یافته‌های پژوهش «عبدالهی و جوکار» (۱۳۸۸) نیز نشان داد بین شکل نوشتاری واژه و ابزار جستجو رابطه وجود دارد؛ بدین معنا که به کار بردن یک شکل خاص از کلیدواژه و استفاده از یک ابزار جستجوی خاص، بر بازاریابی اطلاعات تأثیرگذار خواهد بود. به عنوان مثال، جستجوی شکل پیوسته نویسی مصداق «سوراخکاری» در پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی» می‌تواند به بازاریابی مدارکی بینجامد که تنها شکل پیوسته نویسی مصداق در آن روی داده است و در مقابل، در پایگاه‌های اطلاعاتی دیگر همچون «بانک اطلاعات نشریات کشور» جستجو با همان شکل پیوسته مصداق «سوراخکاری» می‌تواند به بازاریابی مدارک دارای دو شکل نگارشی جدا و پیوسته منتهی شود. در کنار یافته‌های این دو پژوهش، یافته‌های «گل تاجی و بذرگر» (۱۳۸۹)، «آخشیک و فتاحی» (۱۳۹۱)، «هماوندی و دیگران» (۱۳۹۷) نیز به طور کلی نشان داد پایگاه‌های اطلاعات علمی فارسی به مسائل ریخت‌شناسی زبان فارسی از جمله پیوسته نویسی و جدانویسی توجه چندانی نشان نداده‌اند.

افزون بر موارد بالا، نتایج به دست آمده از توزیع فراوانی قابلیت پایگاه‌های اطلاعات علمی فارسی از نظر گسترش جستجو و بازاریابی مدارک به صورت پیوسته نویسی و

جدانویسی برحسب قاعده‌های دهگانه فرهنگستان زبان و ادب فارسی نشان داد، پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» قادر است برخی مصداق‌های انتخابی مربوط به قاعده‌های ۱، ۲، ۵ و ۹ با هریک از شکل‌های نگارشی مصداق‌ها اعم از پیوسته‌نویسی و جدانویسی مدارک مورد نظر را بازیابی کند. به عنوان مثال، در قاعده ۹ که مربوط به ترکیب‌های دارای پسوند است، با جستجوی مصداق‌ها «حسگر» و «نوسانگر»، پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» قادر است با هر دو حالت جدانویسی و پیوسته‌نویسی، مدارک مورد نظر را بدون توجه به شکل نگارشی دوباره بازیابی کند. از دلایل این امر می‌تواند این نکته باشد که پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» احتمالاً از روش‌های خاصی جهت بهنجارسازی شکل‌های مختلف یک مصداق در برخی قاعده‌های نگارشی استفاده می‌کند.^۱

در مقابل، پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» در قاعده‌های بیشتری (۱، ۲، ۳، ۸، ۹ و ۱۰)، ولی از نظر توانایی بازیابی مدارک با هریک از شکل‌های نگارشی، در دامنه‌ای محدودتر نسبت به پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» جای داشت؛ بدین ترتیب که در صورت جستجوی مصداق‌ها با حالت جدانویسی در پایگاه اطلاعات علمی فارسی مذکور، علاوه بر بازیابی مدارک شامل شکل نگارشی جدانویسی از مصداق به بازیابی مدارک دارای شکل نگارشی نزدیک نویسی مصداق نیز می‌انجامید. به عنوان مثال، در قاعده ۱ که مربوط به جدانویسی ترکیب‌های طولانی و پردندانه است، با جستجوی شکل جدانویسی مصداق «آب شستگی»، پایگاه اطلاعات علمی فارسی مذکور قادر بود، علاوه بر بازیابی مدارک دارای آن شکل، مدارک دارای شکل نزدیک نویسی مصداق (آب شستگی) را هم بازیابی نماید. این نتایج با یافته‌های به دست آمده از پژوهش «آخشیک و فتاحی» (۱۳۹۱) مغایرت

۱. علی‌رغم مکاتبه‌ای که با تیم پشتیبانی فنی پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» انجام شد، اما تیم مربوطه به دلیل مسائلی رقابتی با سایر پایگاه‌های اطلاعات علمی فارسی حاضر به معرفی روش‌های مورد استفاده نشدند.

دارد. نتایج پژوهش آنان نشان داد پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» به ویژگی‌های پیوسته‌نویسی و جدانویسی توجه نشان نداده است و تنها حالت ثبت‌شده عنوان مدارک به بازیابی دوباره آن می‌انجامد، درحالی‌که یافته پژوهش حاضر نشان داد پایگاه اطلاعات علمی فارسی مورد نظر در مواردی که مدرک ذخیره‌شده در پایگاه اطلاعات علمی فارسی دارای شکل نگارشی نزدیک‌نویسی از مصداق (واژه یا کلیدواژه انتخابی) باشد، در صورت جستجوی مصداق مورد نظر با شکل نگارشی جدانویسی، پایگاه همچنان قادر به بازیابی مدرک مورد نظر است. با وجود این، در ارتباط با رویکرد پایگاه «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» در بهنجارسازی حالت‌های نگارشی یک مصداق می‌توان گفت از آنجا که یکی از مشکلات مهم خط فارسی در باب واژگان مرکب، مسئله «مشخص نبودن مرز دقیق واژگان مرکب» است (مرتضایی، ۱۳۸۰؛ رضایی، ۱۳۸۶؛ آخشیک و فتاحی، ۱۳۹۱)، این راهکار پایگاه اطلاعات علمی فارسی بالا ضمن افزایش جامعیت نتایج بازیابی شده می‌تواند سبب همارایی نادرست واژگان شده و در نتیجه به ریزش کاذب در مدارک بازیابی شده منتهی شود. نتایج به‌دست‌آمده از مقایسه شکل نگارشی مصداق‌ها در عنوان، چکیده و واژگان کلیدی نسخه پی.دی.اف مدارک با اطلاعات ارائه‌شده از همان مدارک در پایگاه‌های اطلاعات علمی فارسی نیز نشان داد، شکل نگارشی بخش قابل توجهی از مصداق‌های مدارک تغییر یافته است؛ بدین ترتیب که در دو پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» و پایگاه اطلاعات علمی «مرکز اطلاعات علمی جهاد دانشگاهی» از مجموع ۱۴۴ مدرک برای هر پایگاه اطلاعات علمی فارسی، به ترتیب شکل نگارشی ۵۶ و ۵۹ مصداق تغییر یافته است و بیشترین تغییرات و عدم یکدستی نیز مربوط به مدارک دارای شکل نگارشی نزدیک‌نویسی است که احتمالاً در زمان درون‌برد^۱ به درون پایگاه اطلاعات علمی فارسی به شکل نگارشی جدا یا پیوسته تغییر یافته است

و در سایر تغییرات نگارشی که از فراوانی کمتری برخوردار است، شکل نگارشی مصداق‌ها از حالت نگارشی «پیوسته به جدا» تغییر یافته است.

در پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» وضعیت کمی متفاوت بود و نسبت به دو پایگاه اطلاعات علمی فارسی دیگر پراکندگی و بی‌نظمی بیشتری در ناهمسانی شیوه نگارشی مصداق‌ها بین فایل پی.دی.اف مدارک و پایگاه اطلاعات علمی فارسی مشاهده شد؛ اما مشابه دو پایگاه اطلاعات علمی فارسی دیگر، بیشتر این تغییرات مربوط به مدارک دارای شکل نگارشی نزدیک نویسی بود که در پایگاه اطلاعات علمی فارسی عمدتاً به شکل نگارشی جدا یا پیوسته تغییر یافته بود. بنابراین، با توجه به اطلاعات و یافته‌های به دست آمده از این بخش از پژوهش می‌توان گفت آنچه از یک مدرک جهت جستجو و بازیابی مدارک در پایگاه‌های اطلاعات علمی فارسی قرار داده می‌شود، رفتار اولیه نگارنده مدارک نیست. این تغییرها و عدم یکدستی در شکل نگارشی مصداق‌ها بین مدارک (عنوان، چکیده و واژگان کلیدی) و همان اطلاعات در پایگاه اطلاعات علمی فارسی، به چند دلیل احتمالی می‌تواند روی داده باشد که توضیح آن در ادامه می‌آید.

مقاله‌های مجله‌های علمی عمدتاً به شکل نسخه ورد^۱ و پی.دی.اف در اختیار پایگاه‌های اطلاعات علمی فارسی قرار می‌گیرد و نمایه‌سازان پایگاه‌های اطلاعاتی نیز به شکل متداول (کپی و درج^۲)، بخش عنوان، چکیده و واژگان کلیدی، هر مقاله را به درون نرم‌افزار مدیریت محتوای پایگاه‌های اطلاعاتی وارد می‌کنند. این شیوه ورود اطلاعات، در صورت عدم پشتیبانی و شناسایی شکل نگارشی نزدیک نویسی، می‌تواند از دلایل احتمالی بروز خطا و در نتیجه تبدیل شدن شکل نزدیک نویسی عمده مصداق‌ها به حالت جدانویسی باشد. علاوه بر این، از آنجاکه بخش‌های چکیده، عنوان و واژگان کلیدی مقاله‌های فاقد فایل ورد، توسط گروه تایپ‌بست هر پایگاه اطلاعات علمی فارسی

1. Word
2. Paste

تایپ می‌شود، از دیگر دلایل احتمالی تفاوت و عدم یکدستی شکل نگارشی مصداق‌ها، همچون تغییر شکل نگارشی مصداق‌ها از حالت «جدا به پیوسته» یا «پیوسته به جدا» بین نسخه پی.دی.اف مدرک و پایگاه اطلاعات علمی فارسی، می‌تواند ناشی از این مسئله باشد. علاوه بر این، در پایگاه اطلاعات علمی فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»، عامل احتمالی دیگری در ایجاد این ناهماهنگی در شکل نگارشی مصداق‌ها دخیل است؛ پایگاه اطلاعات علمی فارسی مذکور برای بخش واژگان کلیدی تعداد معینی کلیدواژه در نظر گرفته است و در صورت نبود کلیدواژه یا کم بودن آن از حد تعیین شده، نمایه‌سازان پایگاه اطلاعات علمی فارسی مذکور، تعدادی کلیدواژه جدید به بخش کلیدواژه‌ها اضافه می‌کنند. این مسئله نیز می‌تواند از دیگر دلایل احتمالی تفاوت و ناهماهنگی شکل نگارشی مصداق‌ها بین نسخه پی.دی.اف مدارک و پایگاه اطلاعات علمی فارسی باشد.

به‌طور خلاصه، بنا بر یافته‌ها و مطالعه‌های پیشین می‌توان این‌گونه جمع‌بندی کرد که هرچند دو پایگاه اطلاعات علمی فارسی «بانک اطلاعات نشریات کشور» و «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» در برخی مصداق‌ها مربوط به قاعده‌های دهگانه به مسئله پیوسته‌نویسی و جدانویسی توجه نشان داده‌اند و در نتیجه قادر به بازیابی مدارک با دیگر شکل‌های نگارشی مصداق‌هاست؛ اما با توجه به اینکه تنها ۲۳ مدرک معادل ۳/۵٪ از مدارک در هنگام جستجو با هریک از شکل‌های نگارشی قابل بازیابی است و در مقابل ۴۰۹ مدرک دیگر معادل ۷/۹۴٪، تنها با همان حالت ثبت شده مصداق‌ها مدارک قابل بازیابی است، می‌توان نتیجه گرفت پایگاه‌های اطلاعات علمی فارسی چالش‌های پیوسته‌نویسی و جدانویسی را به‌منظور بهبود جامعیت نتایج جستجو چندان مورد توجه قرار نداده‌اند. این بی‌توجهی در کنار تغییرهای شکل نگارشی واژگان در زمان ورود اطلاعات مدارک (عنوان، چکیده و واژگان کلیدی) به درون پایگاه‌های اطلاعاتی علمی فارسی، می‌تواند نایکدستی در شکل نگارشی مصداق‌های مدارک را

تشدید کند. در نهایت، کاربران بالقوه این قبیل پایگاه‌های اطلاعاتی در صورت جستجو با یک شکل نگارشی، از دستیابی جامع به مدارک مورد نیازشان بازمی‌مانند.

پیشنهاد‌های پژوهش

پیشنهاد‌های کاربردی

با توجه به نتایج به دست آمده از پژوهش حاضر، پیشنهاد‌های زیر می‌تواند در جهت بهبود یا رفع چالش‌های پیش روی پایگاه‌های اطلاعات علمی فارسی، مؤثر واقع شود. آگاه‌سازی کاربران نسبت به چالش‌های رسم‌الخط فارسی و همچنین چگونگی استفاده از پایگاه‌های اطلاعات علمی فارسی. به عنوان نمونه، «اختصاصی سازی صفحه ورود پایگاه اطلاعات علمی فارسی برای هر کاربر، جهت آگاهی از قلمرو موضوعی کاربران و سنجش گرایش نگارشی کاربران هر قلمرو موضوعی و در نتیجه اعمال آنها در فرایندهای بعدی ذخیره، پردازش و بازیابی مدارک مرتبط»، «نمایش یک پیغام یا یک فیلم آموزشی کوتاه، جهت آشنایی کاربران با راهبردهای جستجو و همچنین وجود مسائل و چالش‌های رسم‌الخط فارسی تأثیرگذار بر بازیابی مدارک در پایگاه اطلاعات علمی فارسی»، «نمایش یک پیام به کاربر در صورت جستجو با یک شکل نگارشی و یا پایین بودن تعداد نتایج بازیابی شده از یک حد معین» و «آگاه‌سازی کاربران نسبت به تکرار جستجو با دیگر شکل‌های نگارشی یک واژه، با قرار دادن متنی پیش فرض در درون جعبه جستجو»؛

برگزاری نشست‌هایی از سوی طراحان پایگاه‌های اطلاعات علمی فارسی. پیشنهاد می‌شود طراحان پایگاه‌های اطلاعات علمی فارسی نشست‌هایی را با هدف ارائه تجربه‌ها و بهره‌مندی از دستاوردهای یکدیگر در زمینه رفع یا بهبود چالش‌های خط و زبان فارسی در محمل‌های الکترونیکی برگزار کنند.

رفع خط‌های احتمالی و اشتباهات انسانی در زمان درون برد مدارک (چکیده، عنوان و واژگان کلیدی) به درون پایگاه‌های اطلاعات علمی فارسی. پیشنهاد می‌شود طراحان و نمایه‌سازان پایگاه‌های اطلاعاتی ضمن توجه به رخدادهای مسئله بالا، نسبت به رفع آن اقدام

کنند. همچنین، پیشنهاد می‌شود آن دست از مدارکی که بخش عنوان، چکیده و واژگان کلیدی آنها از سوی گروه تاپیست پایگاه‌های اطلاعاتی حروف چینی مجدد می‌شود، قبل از انتقال به محیط پایگاه اطلاعات علمی فارسی از نظر غلط‌های املایی و همچنین واژگان دارای چند شکل نگارشی اعم از پیوسته‌نویسی و جدانویسی ویراستاری شوند. به‌کارگیری روش N-gram در الگوریتم‌های نمایه‌سازی پایگاه‌های اطلاعات علمی فارسی. از آنجاکه این نوع نمایه‌سازی نیازی به اطلاعات قبلی درباره مفاهیم یا زبان متن مورد نظر ندارد (دانش و دیگران، ۲۰۱۱) پیشنهاد می‌شود طراحان پایگاه‌های اطلاعات علمی فارسی برای نمایه‌سازی واژگان مرکب یا دارای چند شکل نگارشی نیز از این شیوه نمایه‌سازی استفاده کنند.

امکان‌سنجی به‌کارگیری قاعده‌های دهگانه در الگوریتم‌های نمایه‌سازی یا بازیابی مدارک. با به‌کارگیری این قاعده‌ها در سامانه‌های بازیابی می‌توان روند نمایه‌سازی خودکار را قاعده‌مندتر و به‌تبع آن روند بازیابی مدارک را بهبود بخشید. به‌عنوان مثال، با به‌کارگیری قاعده دودر الگوریتم‌های نمایه‌سازی می‌توان به مدرک شامل واژه «آب بندی» دیگر شکل‌های نگارشی آن، از جمله «آبندی» و «آب‌بندی» را هم به‌طور خودکار اضافه کرد و به این شکل باعث گسترش نمایه در پایگاه اطلاعات علمی فارسی شد.

به‌کارگیری قابلیت «پیشنهاد واژگان جستجو» در پایگاه‌های اطلاعات علمی فارسی. این راهکار در حال حاضر توسط گوگل و همچنین برخی پایگاه‌های اطلاعاتی فارسی مورد استفاده قرار می‌گیرد و تا اندازه‌ای توانسته است به نزدیک‌تر کردن رفتار نگارشی کاربران با نگارندگان پایگاه‌های اطلاعاتی و در نتیجه بهبود جامعیت نتایج کمک کند.

پیشنهاد‌های پژوهشی

- پیشنهاد می‌شود در پژوهشی مشابه با استفاده از روش استفاده‌شده در این پژوهش، به بررسی موتورهای کاوش (فارسی / غیرفارسی) از نظر میزان توجه به چالش‌های پیوسته‌نویسی و جدانویسی در زبان و خط فارسی انجام گیرد؛

- با توجه به تشابه و نزدیکی خط فارسی و عربی، پیشنهاد می‌شود در پژوهشی به بررسی پایگاه‌های اطلاعاتی عربی پرداخته شود. با انجام این پژوهش می‌توان به مقایسه میزان توجه، استفاده از راهکارهای احتمالی و همچنین الگوبرداری از پایگاه‌های اطلاعاتی عربی پرداخت.

منابع

- آخشیک، سمیه سادات و فتاحی، رحمت‌الله (۱۳۹۱). «تحلیل چالش‌های پیوسته نویسی و جدا نویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی»، کتابداری و اطلاع‌رسانی، (۳)، ۹-۳۰.
- اکبری‌نژاد، سعید (۱۳۷۶). «فاصله خالی میان واژه‌ها در ذخیره و بازیابی رایانه ای اطلاعات». مطالعات ملی کتابداری و سازماندهی اطلاعات، ۱ (۸)، ۴۹-۵۶.
- حری، عباس (۱۳۷۲). «کامپیوتر و رسم الخط فارسی». تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی، ۱ (۳)، ۱۱-۶.
- دلاور، علی (۱۳۹۰). احتمالات و آمار کاربردی در روانشناسی و علوم تربیتی. تهران: رشد.
- راثی، محمدصابر (۱۳۸۵). «مشکلات جست و جو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی: کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر»، مطالعات ملی کتابداری و سازماندهی اطلاعات، ۱ (۱۷)، ۱۷۹-۱۹۶.
- رضائی، مریم (۱۳۸۶). بررسی مشکلات رسم الخط فارسی در بازیابی منابع از وب از دیدگاه کاربران و ارائه راه حل برای این مشکلات (پایان‌نامه کارشناسی ارشد)، دانشکده دانشگاه الزهرا، دانشکده علوم تربیتی و روانشناسی، تهران.
- سرمد، زهره؛ بازگان، عباس و حجازی، الهه (۱۳۹۵). روش‌های تحقیق در علوم رفتاری، تهران: آگه، ۱۳۹۵.
- عبدالهی، محمدصادق و جوکار، عبدالرسول (۱۳۸۸). «چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب». مطالعات تربیتی و روان‌شناسی، ۱۰ (۲)، ۱۸۰-۲۰۱.
- فرهنگستان زبان و ادب فارسی. (۱۳۸۹). دستور خط فارسی. تهران: فرهنگستان زبان و ادب فارسی، گروه نشر آثار.
- قدس‌نیا، پدram؛ زارع بیدکی، علی محمد و یزدانی، ناصر (۱۳۸۶). «بررسی آماری تأثیر برخی از مشکلات زبان فارسی بر جامعیت نتایج جستجو در موتورهای جستجو»، مقاله ارائه شده در سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران.
- کاشفی، امید؛ نصری، میترا و کنعانی، کامیار (۱۳۸۹). خطایابی/املایی خودکار در زبان فارسی، (شورای عالی اطلاع‌رسانی). تهران: شورای عالی اطلاع‌رسانی، دبیرخانه.

- گل تاجی، مرضیه و بذرگر، سعیده (۱۳۸۹). بررسی مشکلات ریخت شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه ای اطلاع‌رسانی علوم و فناوری پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی. *کتابداری و اطلاع‌رسانی*، (۵۰)، ۱۹۹-۲۲۲.
- مرتضایی، لیلا (۱۳۸۰). مسائل زبان و خط فارسی در ذخیره سازی و بازیابی اطلاعات، *پژوهشنامه پردازش و مدیریت اطلاعات*، (۲-۱)، ۲۴-۲۹.
- ولیخانی، احمد و سرافراز، مهدی رضا (۱۳۹۴). روش تحقیق در روانشناسی. تهران: بینش نو.
- هماوندی، هدی؛ نوروزی، یعقوب و حسینی بهشتی، ملوک السادات (۱۳۹۷). «بررسی مشکلات جستجو و بازیابی اطلاعات در پایگاه‌های اطلاعاتی از جنبه ویژگی‌های نگارشی زبان فارسی»، *پژوهشنامه پردازش و مدیریت اطلاعات*، (۳) ۳۳، ۱۰۹۹-۱۱۲۲.
- یوسفی، احمد (۱۳۷۶). «ریزش کاذب در ذخیره و بازیابی اطلاعات»، *پژوهشنامه پردازش و مدیریت اطلاعات*، (۱) ۱۳، ۹-۱.

- Danesh, M.; Minaei, B.; & Kashefi, O. (2011). Challenging Massive Information Retrieval in Persian. *International Journal of Information and Education Technology*, 1(3), 212.
- Hammo, B. H. (2009). Towards enhancing retrieval effectiveness of search engines for diacritized Arabic documents. *Information Retrieval*, 12 (3), 300-323.
- Lazarinis, F. (2007). At the sharp END evaluating the searching capabilities of commerce websites in a non-English language A Greek case study. *Online Information Review*, 31 (6), 881-891.
- Lewandowski, D. (2008). Problems with the use of Web search engines to find results in foreign languages. *Online Information Review*, 32 (4), 668-672.
- Zhang, J.; & Lin, S. (2007). Multiple language supports in search engines. *Online Information Review*, 31 (4), 532-516.