

بررسی مشکلات جستجو و بازیابی اطلاعات در پایگاه‌های اطلاعاتی از جنبه ویژگی‌های نگارشی زبان فارسی

هدی هماوندی*

دانشجوی دکتری علم اطلاعات و دانش‌شناسی

دانشگاه تهران

یعقوب نوروزی

دکتری علم اطلاعات و دانش‌شناسی

دانشیار دانشگاه قم

ملوک السادات حسینی بهشتی

دکتری تخصصی زبان‌شناسی عمومی

استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

پذیرش: ۹۵/۱۱/۱۱

دریافت: ۹۵/۰۴/۰۴

چکیده:

هدف: این پژوهش با هدف 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 انجام شد.

روش: این پژوهش به روش پیمایشی-تحلیلی و با استفاده از شیوه مشاهده مستقیم انجام گرفت. پس از مرور پژوهش‌های مرتبط، کلیدواژه‌های کاوش در قالب یک سیاهه شکل گرفت. هریک از این کلیدواژه‌ها در پایگاه‌های اطلاعاتی مورد مطالعه شامل پژوهشگاه علوم و فناوری اطلاعات ایران، پایگاه استنادی علوم جهان اسلام، پایگاه مجلات تخصصی نور و پایگاه اطلاعات علمی جهاد دانشگاهی جستجو و تعداد نتایج بازیابی شده ثبت گردید. سپس به بررسی میزان انطباق پایگاه‌های اطلاعاتی با این ویژگی‌ها پرداخته شد.

یافته‌ها: برخی ویژگی‌های نوشتاری و معنایی زبان فارسی سبب بروز مشکلاتی در بازیابی اطلاعات از پایگاه‌های اطلاعاتی منتخب می‌شوند. مواردی مانند پیوسته‌نویسی و جدانویسی واژگان مشتق، مرکب و مشتق-مرکب، گوناگونی جمع‌ها، واژگان دخیل و معادل آنها در بخش نوشتاری و چندمعنایی، هم‌نامی و در بخش معنایی از این دست ویژگی‌ها هستند. فقدان پوشش مناسب ویژگی‌های یاد شده در مراحل ذخیره‌سازی و پردازش و عدم آگاه نمودن کاربر از آن، جهت اصلاح فرایند کاوش در مرحله بازیابی اطلاعات در پایگاه‌های اطلاعاتی مورد پژوهش، اثرات نامطلوبی بر فرایند کاوش و بازیابی دارد.

نتیجه‌گیری: یافته‌ها نشان داد که پایگاه‌های اطلاعاتی فارسی نسبت به ویژگی‌های نوشتاری و معنایی زبان فارسی توجه کافی نداشته و بسیاری از ویژگی‌های آنرا در مراحل ذخیره‌سازی و پردازش اطلاعات نادیده

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۲۲۵۱-۸۲۲۳
شاپا(الکترونیکی) ۲۲۵۱-۸۲۳۱
نمایه در SCOPUS، LISTA و ISC
http://jlist.irandoc.ac.ir
دوره XX | شماره X | صص XX-XX
۱۳XX X

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:

درون متنی:

(هماوندی، نوروزی و حسینی بهشتی، ۱۳۹۵)

در فهرست منابع:

هماوندی، هدی، یعقوب، نوروزی، حسینی بهشتی، ملوک سادات. ۱۳۹۵. بررسی مشکلات جستجو و بازیابی اطلاعات در پایگاه‌های اطلاعاتی از جنبه ویژگی‌های نگارشی زبان فارسی. پژوهشنامه پردازش و مدیریت اطلاعات. <http://jipm.irandoc.ac.ir> (دسترسی در روز/ماه/سال)

می‌گیرند. با توجه به تأثیر این ویژگی‌ها در تعامل کاربران با پایگاه‌های اطلاعاتی، احتیاج کاربران فارسی‌زبان به ابزارهای کاوش بومی و پایگاه‌های اطلاعاتی که مبتنی بر ویژگی‌های زبانی خودشان طراحی شده باشد، بیش از پیش احساس می‌شود. پژوهش حاضر با بررسی میزان توانایی پایگاه‌های اطلاعاتی فارسی‌زبان در پوشش برخی ویژگی‌های این زبان که در فرایند جستجو و بازیابی تأثیر قابل توجهی دارند، نقاط ضعف و قوت این پایگاه‌ها را مشخص نموده و نتایج آن می‌تواند در جهت بهبود و اصلاح عملکرد پایگاه‌های مذکور مورد استفاده قرار گیرد.

کلیدواژه‌ها: بازیابی اطلاعات؛ پایگاه‌های اطلاعاتی؛ زبان فارسی؛ ویژگی‌های نگارشی

*هدی هموندی h.homavandi@ut.ac.ir

۱. مقدمه

رایانه‌ها و بستر شبکه جهانی وب بدون اغراق از عمده‌ترین و پر استفاده‌ترین ابزارهای دسترسی به اطلاعات در دنیای امروز هستند. این موضوع سبب بروز چالش‌ها و فرصت‌هایی شده است. یکی از این چالش‌ها، تنوع زبان‌های مورد استفاده و نحوه پشتیبانی ابزارهای دسترسی به اطلاعات از آنها است. مسئله استفاده از زبان طبیعی در محیط‌های رایانه‌ای از دیدگاه رویتر (۲۰۰۶) ابعاد گوناگونی دارد که بحث بازیابی اطلاعات معمولاً به‌عنوان دغدغه‌ای همیشگی در آن طرح می‌شود. تعامل کاربران با اطلاعات و شبکه‌های ارتباطی و اطلاعاتی نظیر اینترنت، وب، پایگاه‌های اطلاعاتی و موتورهای کاوش همواره با میانجی زبان است. آنها از این طریق مفاهیم مورد نظر خود را بیان و اطلاعات را کاوش و بازیابی می‌کنند. مارچونینی (۲۰۰۸) عقیده دارد که امروزه یکی از اساسی‌ترین دغدغه‌ها در حوزه علم اطلاعات توضیح چگونگی تعامل انسان با آنچه خود ساخته هست که ابعاد و مباحث گوناگونی از جمله زبان را در بر می‌گیرد. همین تنوع و تکثر است که آنها را به یک حوزه میان رشته‌ای تبدیل کرده است. زره‌ساز و فتاحی (۱۳۸۵) نیز اشاره می‌کنند که در سال‌های اخیر این حوزه با تأثیرپذیری از حوزه‌هایی مانند روان‌شناسی، علوم رایانه، علم اطلاعات، جامعه‌شناسی، و سایر حوزه‌های مشابه گسترش بسیاری یافته و متون و تحقیقات فراوانی نیز در این زمینه پدید آمده است. با توجه به این رویکرد، دور از ذهن نیست اگر یکی از موارد مؤثر ذکر شده را که رایانه‌ها به‌عنوان ابزارهای اطلاعاتی همواره در مواجهه با پیچیدگی‌های آن مشکلاتی دارند، زبان طبیعی دانست. البته در این بیان منظور از زبان طبیعی، زبانی است که انسان‌ها معمولاً در محاورات یا نوشته‌های خود از آن استفاده می‌کنند و به ناچار ممکن است چالش‌هایی را در این زمینه ایجاد کند.

آمارهای مربوط به سال ۲۰۱۵ در مورد استفاده از اینترنت بر اساس زبان حاکی از آن است که حدود ۶۲/۴ درصد کاربران انگلیسی‌زبان و ۳۷/۶ درصد غیرانگلیسی‌زبان هستند (وب‌سایت آمارهای جهانی اینترنت، ۲۰۱۵). این امر نشان از رشد روزافزون کاربران غیرانگلیسی‌زبان داشته و لزوم توجه به نیازهای زبانی و حل مسائل مربوط به آنها را (با توجه به اینکه زبان غالب در شبکه

وب و ابزارهای کاوش انگلیسی است) یادآور می‌شود. کاربران فارسی زبان نیز با ویژگی‌های خاص زبانی خود در زمره دومین گروه هستند. از همین روست که در سال‌های اخیر بسیاری از تحقیقات بین رشته‌ای به بررسی تأثیر مسائل زبانی در تعامل کاربران و محیط وب معطوف شده‌اند. تحقیقاتی که هریک از دیدگاهی بحث زبان در کاوش و بازیابی اطلاعات را در انواع رسانه‌های اطلاعاتی اعم از پایگاه‌های تحت وب، موتورهای کاوش، وب‌سایت‌های تجاری، کتابخانه‌ای و... مورد کنکاش قرار داده و مشکلات ناشی از آنرا بررسی نموده‌اند. عدم توجه به ویژگی‌های تأثیرگذار زبان فارسی از جمله ویژگی‌های نگارشی در مراحل ذخیره‌سازی و پردازش، جستجو و بازیابی اطلاعات از پایگاه‌های اطلاعاتی موجب ایجاد چالش‌ها و موانعی پیش روی کاربران فارسی زبان در دستیابی به اطلاعات مورد نیازشان شده است. به عنوان مثال بازیابی نتایج متفاوت برای صورت‌های مختلف نوشتاری کلیدواژه‌های "گلابگیری"، "گلاب گیری" به عنوان واژگان مشتق-مرکب و یا "انفولانزا، آنفلوآنزا، آنفولانزا" به عنوان صورت‌های مختلف ضبط واژگان یا بازیابی نتایج متفاوت برای کلیدواژه‌های مترادف "دریای خزر، دریای مازندران و دریای کاسپین" از پیامدهای عدم پوشش مناسب آنها در مرحله نمایه‌سازی و پردازش است. لذا مطالعه پیش رو سعی در تبیین مشکلات عمده نگارشی زبان فارسی اعم از نوشتاری و معنایی در تعامل انسان و محیط‌های کاوش وب مانند پایگاه‌های اطلاعاتی دارد. همچنین پژوهش حاضر تلاش می‌کند تا میزان انطباق و سازگاری نمونه‌هایی از پایگاه‌های اطلاعاتی برگزیده داخلی شامل پایگاه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، پایگاه استنادی علوم جهان اسلام (آی اس سی)، پایگاه مجلات تخصصی نور (نورمگز) و پایگاه اطلاعات علمی جهاد دانشگاهی (سید) را با مؤلفه‌های زبان فارسی از حیث جستجو و بازیابی اطلاعات بررسی نموده و ضمن شناسایی چالش‌ها در بخش نوشتاری و معنایی زبان، مقایسه‌ای بین پایگاه‌های یاد شده نیز انجام دهد تا بدین ترتیب راهکارهایی برای پیشگیری و حل مشکلات یاد شده ارائه نماید.

۲. مروری بر پژوهش‌های مرتبط

تاکنون پژوهش‌های گوناگونی انجام گرفته که هریک به نوعی مسائل زبان فارسی را در تعامل انسان با رایانه، موتورهای کاوش، پایگاه‌های اطلاعاتی مختلف بررسی نموده‌اند. مطالعات یاد شده در اکثر موارد معطوف به ویژگی‌های نوشتاری زبان فارسی در بازیابی اطلاعات متنی بوده‌اند. برخی پژوهش‌های صورت گرفته در این زمینه در جدول ۱ ارائه شده است.

جدول ۱. پیشینه‌های داخلی مربوط به موضوع پژوهش

نویسنده / سال	عنوان	هدف	روش	نتیجه
حری (۱۳۷۲)	کامپیوتر و	تبیین مشکلات	مطالعه	استفاده از سیاست‌گذاری واحد برای

یکسان سازی رسم الخط	موردی	رسم الخط فارسی در ذخیره و بازیابی اطلاعات در محیط رایانه ای	رسم الخط فارسی	
ارزیابی راهکارهای ممکن جهت تطبیق رسم الخط فارسی با محیط های رایانه ای	توصیفی	توصیف ناهماهنگی های میان زبان فارسی و نظام های رایانه ای	مسائل رسم الخط فارسی در رویارویی با فناوری نوین اطلاعاتی	نشاط (۱۳۷۹)
هماهنگی و یکدستی در فاصله گذاری	پیمایشی - توصیفی	تیین و تشریح تاثیر فاصله خالی میان واژه ها در ذخیره و بازیابی اطلاعات	فاصله خالی میان واژه ها در ذخیره و بازیابی رایانه ای اطلاعات	اکبری نژاد (۱۳۷۶)
مسائل زبان و خط فارسی سبب کندی مراحل ذخیره و بازیابی اطلاعات، کاهش نسبت بازیافت اطلاعات و تاثیر منفی بر جامعیت نتیجه یک جستجو می شوند	کتابخانه ای - توصیفی	ارائه نمونه هایی از تجربه های واژه گزینی در ذخیره اطلاعات به منظور تسریع و تسهیل ذخیره و بازیابی اطلاعات به	مسائل زبان و خط فارسی در ذخیره سازی و بازیابی اطلاعات	مرتضائی، ۱۳۸۰

		زبان فارسی		
تفکیک و دسته‌بندی دشواری‌های خط فارسی در مقولاتی مثل نحو و ایهام در نقش‌ها هنگام همنشینی واژگان و...	کتابخانه‌ای	تبیین مشکلات خط فارسی برای پردازش رایانه‌ای	دشواری‌های پردازش رایانه‌ای خط فارسی	اسلامی (۱۳۸۱)
استفاده اغلب کاربران از موتور کاوش گوگل و وجود مشکلات بدلیل عدم توجه به شکل‌های مختلف نوشتاری واژه‌ها	پیمایشی - توصیفی	تبیین مشکلات جستجو و بازیابی اطلاعات در اینترنت، اینترنت از دیدگاه کاربران مرکز اینترنت دانشگاه آزاد واحد شبستر	مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی: کاربران مرکز اینترنت دانشگاه آزادی اسلامی واحد شبستر	رائی ساربانقلی (۱۳۸۵)
تاثیر زیاد استفاده از برجسب‌دهی سلسله مراتبی بر حل مسئله هم‌نگاره‌ها و مشخص کردن مرز گروه‌ها نحوی	توصیفی	بکارگیری استانداردی برای حل مسئله هم‌نگاره‌ها و مشخص کردن مرز گروه‌ها نحوی بعنوان مشکلاتی بر سر راه پردازش متون	بکارگیری یک نظام برجسب‌دهی برای تعبیر و تفسیر یک پیکره متنی زبان فارسی	روح پرور و بی جن خان (۱۳۸۶)

		فارسی		
بررسی مشکلات موجود در رسم الخط رایانه‌ای زبان فارسی در مورد حروفی که در رسم الخط رایانه‌ای دارای چند نوع حرف هستند و رفع مشکلات رسم الخط با ارائه پیشنهاد‌های صحیح به کاربران از طریق الگوریتم خطیاب	پیمایشی - توصیفی	ارائه روشی برای یافتن خطاهای املائی واژگان با استفاده از الگوریتم‌های هوشمند و یادگیرنده	روشی جدید در خطیابی املائی در زبان فارسی	رسولی و بیدگلی (۱۳۸۷)
عدم توجه موتورهای کاوش وب به شیوه‌های نگارش زبان فارسی به منظور بهبود کاوش، و وجود رابطه معناداری بین شکل واژه و نوع ابزار جستجو	پیمایش مقایسه‌ای و اسنادی	بررسی مسائلی که کاوشگران فارسی در کاوش ریخت‌های مختلف یک واژه با آن روبه‌رو هستند	چالشهای شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب	عبداللهی نورعلی و جوکار (۱۳۸۸)
تأثیر زیاد چالش‌های ریختی شناخته شده زبان فارسی بر بازیابی اطلاعات در هر یک از سه پایگاه مورد نظر و عدم پرداختن به حل مسائل ریخت‌شناسی واژگان فارسی به شیوه‌ای جامع و قابل ملاحظه از سوی پایگاه‌های مورد پژوهش	پیمایش مقایسه‌ای	بررسی مشکلات ریخت‌شناسی زبان فارسی در پایگاه‌های اطلاعاتی فارسی	بررسی مشکلات ریخت‌شناسی فارسی در سه پایگاه اطلاعاتی آی‌اس‌اسی، سید، ایرانداک	گل‌تاجی و بدرگر (۱۳۸۹)
عملکرد بهتر پایگاه اطلاعاتی ایرانداک نسبت به پایگاه آی‌اس‌سی در بازیابی عنوان پایان	تحلیل	بررسی وضعیت توجه	تحلیل چالش‌های	آخشیک و

<p>نامه‌ها در حالت‌های مختلف پیوسته و جدا نوشته شده و لزوم تأکید به نگارندگان پایان نامه‌ها، به استفاده از قواعد یکدست ملی به ویژه در نگارش کلمات دو جزئی و مشتق</p>	<p>محتوا</p>	<p>نویسندگان و همچنین پایگاه‌های مورد مطالعه به ویژگی پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی و ارائه راهکارهایی برای حل این مشکلات</p>	<p>پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی</p>	<p>فتاحی (۱۳۹۱)</p>
<p>ذکر بیش از ۴۰ دشواری نگارشی موثر بر بازیابی در رابطه با جستجو و بازیابی اطلاعات فارسی در آثار مورد بررسی و ضرورت توجه به هنجارسازی چند دستی‌های نگارشی و دستوری در طراحی الگوریتم‌های سامانه‌های جستجو و بازیابی فارسی، تدوین استاندارد نگارش فارسی و تجهیز پایگاه‌های اطلاعاتی به اصطلاحنامه و فرهنگ‌های املائی و...</p>	<p>متن پژوهی با رویکرد تحلیل محتوا</p>	<p>بررسی متون و پیشینه‌های موجود به منظور تبیین چالش‌های نگارش فارسی، تأثیر آنها بر اثربخشی بازیابی اطلاعات، و پیشنهادهایی در جهت رفع این دشواری‌ها</p>	<p>مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تاثیرات آنها بر اثر بخشی پردازش خودکار متن و بازیابی اطلاعات</p>	<p>ستوده و هنرجویان (۱۳۹۱)</p>

همان‌طور که در جدول ۱ مشاهده می‌شود، مرور پیشینه‌های مورد مطالعه که روش غالب آنها پیمایش است در مجموع ضمن تبیین و شناسایی بسیاری از ویژگی‌ها و مشکلات نگارشی زبان فارسی، حاکی از آن است که رسم‌الخط فارسی یکی از متغیرهای عمده در ذخیره و بازیابی اطلاعات خصوصاً از موتورهای کاوش و پایگاه‌های اطلاعاتی زبان فارسی است.

در خارج از ایران و پیرامون سایر زبان‌ها نیز مطالعات مشابه گوناگونی انجام شده است که برخی از آنها در جدول ۲ ارائه شده است.

جدول ۲. پیشینه‌های خارجی مربوط به موضوع پژوهش

نویسنده/سال	عنوان	هدف	روش	نتیجه
لازارینیس ^۱ (۲۰۰۸)	بهبود مبتنی بر مفهوم بازیابی تصاویر در وب از طریق ترکیب کلیدواژه‌های مشابه معنایی در زبان یونانی	بررسی عملکرد موتورهای کاوش (گوگل، یاهو و ام اس ان) در ارتباط با سئوالات تک‌زبانه غیرانگلیسی (یونانی) و ارائه ابزار فراکاوش برای اصلاح نقایص و در جهت ایجاد موتورهای جستجوی کارآمد بومی	پیمایشی	تأثیر قابل توجه ریخت‌شناسی کلمات (استفاده از حروف کوچک یا بزرگ، استفاده یا عدم کاربرد علائم تلفظ زبانی، و حتی نقش واژه‌ها) بر بازیابی نتایج (تعداد و ربط تصاویر) و تکیه موتورهای جستجو بر شکل کلیدواژه‌ها به جای تمرکز بر نیاز واقعی کاربران و عدم توجه موتورهای کاوش مورد پژوهش به ویژگی‌های زبان‌شناختی دیگر زبان‌ها
لازارینیس (۲۰۰۷)	قابلیت‌های جستجوی وب سایت‌های الکترونیکی تجاری در مورد زبان‌های غیرانگلیسی (مطالعه)	بررسی قابلیت‌های جستجوی وب سایت‌های الکترونیکی تجاری در مورد زبان یونانی	پیمایشی (مطالعه موردی)	عدم توجه موتورهای جستجوی محلی به ریخت‌شناسی سئوالات که امری مهم و قابل توجه در

^۱-Lazarinis

<p>مورد زبان‌های غیرانگلیسی و غیرلاتین از جمله زبان یونانی است و در نتیجه شکست جستجوی کاربر</p>			<p>موردی یونانی)</p>	
<p>پشتیبانی چندزبانه بهتر موتورهای جستجوی گوگل، EZ2Find و آنلاین لینک^۱ در بین بسیاری از موتورهای جستجوی مجهز به ویژگی پشتیبانی چندزبانه</p>	<p>پیمایش مقایسه‌ای</p>	<p>بررسی ویژگی‌های پشتیبانی چندزبانه بوسیله موتورهای جستجوی شبکه اینترنت</p>	<p>پشتیبانی چندزبانه بوسیله موتورهای جستجو</p>	<p>ژانگ و لین^۱ (۲۰۰۷)</p>
<p>روبرو شدن کاربر با مشکلاتی در موتورهای کاوش گوگل و ام اس ان وقتی نتایج به زبان خاصی محدود می‌شوند، در حالی که هیچ یک از موتورهای کاوش در بازیابی نتایج به زبان صفحه رابط کاربر استفاده شده (زبان انتخابی) با مشکل مواجه نمی‌شوند، عدم تأثیر راه‌برد محدودیت زبانی در بهبود جستجو در همه موارد و اثرگذاری بهتر استفاده از صفحه میانجی به زبان بومی در جستجو و بازیابی</p>	<p>پیمایشی</p>	<p>بررسی توانایی موتورهای جستجوی پُر استفاده و اصلی از جمله گوگل، یاهو، ام اس ان و اسک در تشخیص و تمایز میان مدارک به زبان آلمانی از پیشینه‌هایی با زبان انگلیسی</p>	<p>مشکلات استفاده از موتورهای جستجوی وب برای یافتن نتایج به زبان‌های خارجی</p>	<p>لنداوسکی^۳ (۲۰۰۸)</p>

³- Zhang& Lin

¹- Onlinelink

³-Lewandowski

اطلاعات در برخی از مواقع				
تأثیر گسترش سؤال بر بهبود جستجو و بازیابی متون عربی و افزایش کارآیی موتورهای کاوش با استفاده از ابزارهای پیشرفته پردازش زبان طبیعی	پیمایشی	ارائه یک قالب کاری برای افزایش کارآیی موتورهای کاوش برای متون عربی دارا و فاقد اعراب گذاری از طریق روش‌های گسترش سؤال (کلیدواژه)	افزایش کارآیی موتورهای جستجو برای پیشینه‌های نشانه گذاری شده به زبان عربی	هامو ^۱ (۲۰۰۹)

بر اساس مندرجات جدول ۲، مروری بر پیشینه‌های خارجی که اکثراً با روش پیمایشی انجام شده‌اند، نشان می‌دهد این پژوهش‌ها اغلب به بررسی قابلیت‌ها و کاستی‌های موتورهای کاوش در پوشش زبان‌های غیرانگلیسی و با هدف شناخت و ارائه راهکارهایی جهت اصلاح چالش‌های زبانی پرداخته‌اند. یافته‌های حاصل از آنها نشان می‌دهد که ریخت‌شناسی واژه‌ها و عبارت‌های جستجو شده بر بازیابی نتایج اثر دارد و ابزارهای جستجو به جای تمرکز بر نیاز واقعی کاربران در جهت بهبود فرایند کاوش، بیشتر بر شکل کلیدواژه‌ها تکیه می‌کنند؛ حتی بعضی ابزارهای جستجوی محلی نیز ریخت‌شناسی سؤالات را در نظر نمی‌گیرند؛ بنابراین جستجوی کاربر با شکست مواجه می‌شود.

بررسی پیشینه‌های پژوهش که شناسایی چالش‌های زبانی در ارتباط با جستجو و بازیابی اطلاعات از موتورهای کاوش و پایگاه‌های اطلاعاتی از اهداف عمده آنها بوده و غالباً با روش پیمایشی انجام شده‌اند، در مجموع حاکی از نکاتی است که در ادامه ذکر خواهند شد. با وجود مقبولیت و استفاده زیاد از فضای وب و موتورهای کاوش برای یافتن انواع اطلاعات توسط طیف گسترده‌ای از کاربران و فراهم شدن ابزارهای متنوع با امکانات متفاوت برای جستجو در محیط وب، ویژگی‌ها و مشکلات زبانی همچنان به‌عنوان عاملی مهم در بحث ذخیره و بازیابی اطلاعات مطرح هستند که مطالعات بسیاری به آنها پرداخته و راهکارهایی نیز ارائه شده است. اما در مورد زبان فارسی به دلیل ماهیت و خصوصیات منحصر به فرد آن، نحوه تعامل فارسی‌زبانان با محیط وب و تولید روزافزون صفحات وب و وبلاگ‌های فارسی‌زبان که تا حدودی به آنها اشاره شد، هنوز جای تحقیق و کار بسیار است. همچنین با توجه به آنچه بیان شد، پیشینه‌های داخلی اغلب به ویژگی‌های ریخت‌شناسی زبان فارسی تأکید داشته و ویژگی‌های معنایی آن را کمتر مورد بررسی

^۱-Hammo

قرار داده‌اند. علاوه بر این، تاکنون مطالعه جامعی که پایگاه‌های بررسی شده در پژوهش حاضر را که طیف گسترده‌ای از کاربران از آنها استفاده می‌کنند را همزمان مطالعه کند، یافت نشد. در این میان تنها تعدادی از پژوهش‌ها به توصیف این ویژگی‌ها پرداخته و مشکلات احتمالی رسم‌الخط فارسی را در محیط‌های رایانه‌ای مورد بحث قرار داده‌اند. در صورتی که، مطالعه حاضر با استخراج ویژگی‌ها ریخت‌شناسی و معنایی زبان فارسی، ضمن گردآوری مجموعه‌ای جامع از این ویژگی‌ها، قصد دارد تا با جستجو در محیط پایگاه‌های اطلاعاتی فارسی‌زبان، بر مبنای نتایج بازیابی شده، بصورت عینی مسائل زبانی عمده و اثرگذار در این حوزه را که کاربر با آنها مواجه است را شناسایی کند. بعلاوه، انجام مطالعاتی از این دست می‌تواند از طریق تشخیص نقاط ضعف و قوت پایگاه‌ها (به لحاظ پوشش ویژگی‌های زبانی) زمینه‌ساز بهبود طراحی و اصلاح پایگاه‌های اطلاعاتی و حتی موتورهای جستجوی بومی و ملی باشد.

پرسش پژوهش

میزان انطباق و توجه به ویژگی‌های نگارشی زبان فارسی در پایگاه‌های اطلاعاتی فارسی چگونه است؟

۳. روش‌شناسی پژوهش

این پژوهش به روش پیمایشی-تحلیلی و با استفاده از شیوه مشاهده مستقیم انجام گرفت. بدین منظور، پس از بررسی منابع مرتبط و بر مبنای پیشینه‌های فارسی پژوهش، آن دست از ویژگی‌های زبان فارسی که باعث ایجاد مسائلی در کاوش و بازیابی اطلاعات هستند دسته‌بندی و نسبت به تهیه جداول ۳ و ۴ اقدام شد. جدول ۳ حاوی دسته‌بندی از مشکلات نوشتاری است که بر اساس پیشینه پژوهش باید در ذخیره‌سازی، پردازش و بازیابی منابع مورد توجه قرار گیرند. جدول ۴ نیز طبقه‌بندی از ویژگی‌های معنایی زبان فارسی را نشان می‌دهد که بر بازیابی اطلاعات به این زبان اثرگذار هستند. هر یک از ویژگی‌های ذکر شده در جداول نیز شامل مصادیقی (کلیدواژه‌هایی) است که هر کدام بازنمون چالش‌های نوشتاری یا معنایی زبان فارسی در جستجو و بازیابی اطلاعات در محیط‌های رایانه‌ای هستند. از این کلیدواژه‌ها به منزله سیاهه‌ای برای بررسی و شناسایی مشکلات عمده ناشی از ویژگی‌های نوشتاری و معنایی زبان فارسی در تعامل کاربران فارسی‌زبان با پایگاه‌های اطلاعاتی و در ارتباط با کاوش و بازیابی اطلاعات و نیز چگونگی ایجاد این مسائل استفاده شده است. به این گونه که برای هر یک از ویژگی‌های نوشتاری و معنایی زبان فارسی واژه‌ای انتخاب شد تا به‌عنوان کلیدواژه کاوش، مبنای قرار گیرد و توسط پژوهشگران در پایگاه‌های مورد مطالعه جستجو و نتایج بازیابی شده حاصل از آن بررسی و ثبت شود. روایی سیاهه مذکور که شامل کلیدواژه‌های موجود در جداول ۵ و ۶ است نیز از با مشورت شش نفر از

اساتید علم اطلاعات و دانش‌شناسی و زبان و ادبیات فارسی مورد تأیید قرار گرفت. به‌عنوان نمونه برای ویژگی "همنامی" از واژه "توپ" جهت آزمون مشکلات پیش آمده در کاوش و بازیابی اطلاعات در پایگاه اطلاعاتی مورد نظر، در رابطه با این ویژگی استفاده شد. به‌منظور گردآوری داده‌ها، هریک از کلیدواژه‌های موجود در سیاهه توسط پژوهشگران به تفکیک، وارد بخش جستجوی پایگاه‌های اطلاعاتی مورد مطالعه شامل ایرانداک (که کاوش در آن بالاخص در زمینه جمع‌آوری پیشینه در موضوعات مختلف پژوهشی بسیار مورد توجه محققان و خصوصاً دانشجویان و اساتید است)، آی اس سی (پایگاه دیگری که مورد توجه پژوهشگران و دانشجویان است)، نورمگز (که این پایگاه نیز خصوصاً در حوزه علوم انسانی و اسلامی مراجعان زیادی دارد) و سید (این پایگاه نیز بدلیل محتوایی که دارد مورد مراجعه بسیاری پژوهشگران و دانشجویان است) شد و نتایج حاصل ثبت گردید. در بخش مسائل نوشتاری صورت‌های مختلف متصور برای هر واژه درج و تعداد کل نتایج بازیابی شده ثبت شد. در بخش مسائل معنایی نیز به همین ترتیب عمل شد با این تفاوت که در مورد کلمات فاقد صورت‌های مختلف نوشتاری و دارای معانی گوناگون، مانند واژه‌های توپ و شور؛ در میان بیست نتیجه نخست بازیابی شده، تعداد نتایجی که حاوی معانی مختلف کلیدواژه جستجو شده بودند، شمارش و ثبت شد.

۴. یافته‌های پژوهش

برای پاسخگویی به پرسش پژوهش، همان‌طور که ذکر شد ابتدا با مرور متون و با توجه به پژوهش‌های انجام شده و بر اساس مطالعه هم‌اوندی (۱۳۹۲) ویژگی‌های نوشتاری و معنایی عمده زبان فارسی که در روند بازیابی اطلاعات توسط کاربران به‌ویژه در پایگاه‌های اطلاعاتی چالش ایجاد می‌کنند استخراج و پس از تغییراتی متناسب با محیط پایگاه‌ها به‌طور خلاصه در جدول شماره ۳ و ۴ ارائه شد.

جدول ۳. ویژگی‌های نوشتاری زبان فارسی مؤثر در بازیابی اطلاعات در محیط‌های اطلاعاتی

مثال	مسئله نوشتاری
انفولانزا، آنفلوآنزا، آنفلوانزا/ پتاسیم، پتاسیوم/ امریکا، آمریکا	نحوه ضبط واژگان لاتین
ایمیل، رایانامه/ سیستم، نظام، سامانه	واژگان دخیل و انواع معادل آنها
پرستش گاه، پرستشگاه	پیوسته‌نویسی و جدا نویسی انواع واژگان
مشتق	

^۱ - ۱۶ بهمن ماه ۱۳۹۴ (لازم است تاریخ جستجو و بازیابی نتایج، بدلیل تغییرات مداوم محتوای پایگاه‌های اطلاعاتی مد نظر قرار گیرد)

مرکب	جوانمرد، جوان‌مرد / بزرگسال، بزرگ سال
مشتق مرکب	گلابگیری، گلاب‌گیری / فن‌آوری، فناوری
علائم جمع	دانشگاهها، دانشگاه‌ها
انواع جمع‌ها (جمع‌های فارسی و مکسر)	حوادث، حادثه‌ها / اساتید، استادان، استادها
طریقه نگارش الف مقصوره	موسی، موسا / مصلی، مصللا
استفاده یا عدم استفاده از اعراب گذاری و سایر علائم (واژه‌هایی با شکل نوشتاری یکسان و تلفظ متفاوت)	مسکن (مُسکن، مَسکن) / ملک (مَلک، مُلک، مَلک، مَلِک)
نحوه نگارش همزه میانی و پایانی کلمات با کرسی واو، دندانه، الف و بدون کرسی	جبرئیل، جبرئیل / مؤذن، مؤذن / املاء، املاء / مسأله، مسأله، مسئله
استفاده و عدم استفاده از علامت تشدید	محمد، محمد / معلم، معلم
کسره اضافه و بدل‌های آن	اهدا کتاب، اهداء کتاب، اهدای کتاب
واژه‌های دو املائی (واژه‌هایی با واج یا آوای مشترک و شکل نوشتاری متفاوت)	آزوقه، آذوقه / تهران، طهران
جابه‌جایی ی و همزه در کلمات فارسی	پایین، پائین / آیین، آئین
نحوه نگارش ه غیر ملفوظ و ی میانجی	زلزله بم، زلزله ی بم، زلزله بم
استفاده از زبان محاوره (شکل عامیانه)	خونه، خانه / زموئه، زمانه
کاربرد و حذف مد در کلمات فارسی	فناوری، فناوری / دستاورد، دستاورد

با دقت در موارد ذکر شده در جدول ۳ می‌توان دریافت که، بسیاری از ویژگی‌های ریخت‌شناسی و نوشتاری زبان فارسی ریشه در دو یا چندگانه نویسی یک واژه دارند که همین امر بدلیل ایجاد ناهماهنگی سبب بروز مسائلی در بازیابی اطلاعات می‌شود. به‌عنوان مثال، جستجوی صورت‌های مختلف نوشتاری واژه "فناوری" سبب بازیابی نتایج متفاوت می‌شود. همچنین بعضی ویژگی‌های معنایی زبان فارسی هم در این زمره هستند. مرور پیشینه‌ها نیز حاکی از این است که مواردی مانند

همنامی یا واژه‌های یکسان با معانی متفاوت و واژگان هم‌نویسه با معانی متفاوت، چندمعنایی^۱، هم‌معنایی و مترادف از این قبیل هستند. همان‌گونه که (حسینی بهشتی، ۱۳۸۲) نیز عقیده دارد، در این بین چندمعنایی و مترادف دو مشکل عمده در ارتباطات معنایی واژگان هستند.

مسئله معنایی	مثال
هم‌نامی یا واژه‌های یکسان با معانی متفاوت (واژگان مشترک لفظی/ هم‌آوا و هم‌نویسه)	شیر (نام یک حیوان، ماده لبنی، شیر آب)/ توپ (توپ بازی، توپ جنگی، واحد شمارش پارچه، معنای عوامانه)
واژگان هم‌نویسه با معانی متفاوت	شور (شور بودن طعم، شور و شوق)
چندمعنایی	روان (روح و روان، جاری)/ قلب (عضو بدن، ضمیر و خاطر، وارونه بودن، مرکز)
هم‌معنایی و مترادف	دریای خزر، دریای کاسپین، دریای مازندران

جدول ۴. ویژگی‌های معنایی زبان فارسی مؤثر در بازیابی اطلاعات در محیط‌های اطلاعاتی

ویژگی‌های معنایی ذکر شده در جدول شماره ۴ که در پژوهش‌های این حوزه کمتر مورد توجه قرار گرفته‌اند، اغلب با ایجاد تعدد معنایی سهم عمده‌ای از مشکلات یاد شده در حوزه جستجو و بازیابی اطلاعات را به خود اختصاص می‌دهند. به‌عنوان مثال، کاربرد با جستجوی کلیدواژه "توپ" که دارای ویژگی هم‌نامی است ممکن است با نتایجی حاوی چندین معنی مواجه شده و این کاوش به بازیابی نتایج نامربوط منجر شود و یا اینکه به نتایجی حاوی چند مورد از معانی دست یابد اما در بین آنها نتایجی با معنی مورد نظر وی بازیابی نشود.

پرسش پژوهش: میزان انطباق و توجه به ویژگی‌های نگارشی زبان فارسی در پایگاه‌های اطلاعاتی فارسی چگونه است؟

به‌منظور پاسخ به پرسش پژوهش، شکل‌های گوناگون این کلیدواژه‌ها توسط پژوهشگران در قسمت جستجوی پایگاه‌های ایرانداک، آی اس سی، نورمگز، و سید وارد و سپس تعداد نتایج بازیابی شده برای هر کلیدواژه ثبت و در جدول شماره ۵ ارائه شده‌اند.

^۱ - چندمعنایی تنها مختص زبان فارسی نیست و از ویژگی‌های سایر زبان‌ها نیز می‌باشد که در اینجا اختصاصاً در مورد زبان فارسی بررسی شده است.

جدول ۵. نتایج بازیابی شده برای کلیدواژه‌های منتخب با صورت‌های گوناگون
نوشتاری در پایگاه‌های مورد مطالعه

مجموع نتایج بازیابی شده در پایگاه‌های اطلاعاتی				کلیدواژه‌های مورد جستجو
سید	نورمگز	آی اس سی	ایراندک	
۱	۲۱	۱	۷	پتاسیوم
۴۲۱	۶۵۹	۱۶۶۳	۵۷۳۶	پتاسیم
۵۵۳	۳۹۳۱	۳۶۷۳	۴۵۲۸	سامانه
۳۱۳۲	۵۶۳۸۶	۲۱۶۳۲	۷۳۳۱۶	سیستم
۱۷۲۴	۱۴۱۷۲۶	۱۴۳۱۶	۲۲۲۴۱	نظام
۰	۱۰۷۸	۶	۱۲	پرستشگاه
۰	۸۰۰۳۰	۱۰	۲۷	پرستش گاه
۹۳	۱۱۹۸۱	۳۳۰	۸۱۰	بزرگسال
۱۹	۳۰۱۳۵۵	۱۱۷۵	۳۷۸۱	بزرگ سال
۱۰۲۶	۱۸۹۶۰	۱۰۷۴۹	۱۸۹۳۰	فناوری
۲۰۳	۱۱۰۱۷۱	۱۹۱۸	۳۷۷۲	فن آوری
۲۴	۱۴۳۲۷۵	۵۰۹	۱۴۸۶	دانشگاهها
۲۹۵	۱۸۲۱۹۹	۹۱۹۹	۵۸۸۰	دانشگاه ها
۲۲۱	۱۰۳۰۷۷	۱۳۶۵	۳۷۳۶	اساتید
۹۰	۱۰۳۰۷۷	۹۰۸	۸۶۳	استادان
۰	۱۰۳۰۷۷	۱	۳	استاد ها
۰	۱۴۸	۳	۵	موسا

۵۴	۴۳۹۳۰	۶۰۹	۳۷۰۲	موسی
۳۲۴	۲۵۵۱۵	۲۱۳۷	۵۵۹۵	مسکن
۳۲۴	۲۵۵۱۵	۰	۱۷۲	مُسکن
۳۲۴	۲۵۵۱۵	۰	۴	مَسکن
۸۵۶	۸۲۳۶۸	۴۶۷۷	۵۸۵۶	مسأله
۸۵۶	۸۲۳۶۸	۴۶۷۷	۱۱۱۳۴	مساله
۱۹	۱۷۷۶۲۲	۱۳۷۹	۲۱۳۵۴	مسئله
۱۳۷	۶۱۲۷۲	۱۴۹۰	۷۵۸۱	معلم
۱۳۷	۶۱۲۷۲	۲۲	۷۲	معلم
۰	۲۶۹۸۲۲	۵	۷	اهدای کتاب
۰	۲۶۸۲۰۴	۱	۵	اهدای کتاب
۰	۲۶۹۸۲۲	۵	۶	اهدای کتاب
۶۵۱۵	۱۴۲۸۳۷	۱۰	۸۱۸۳۴	تهران
۱۰	۱۱۲۳۳	۷۹	۶۷	طهران
۸۸	۲۲۸۹۵	۲۲۷	۳۰۳	زلزله بم
۸۹	۲۲۸۹۷	۱	۳۲	زلزله ی بم
۸۸	۲۲۸۹۵	۰	۰	زلزله بم
۴۴۲	۹۲۹۷۹	۴۰۷۲	۶۵۷۱	خانه
۰	۲۰۹۲	۸	۱۰	خونه
۸	۲۸۸۶۲	۴۸۴	۸۹۷	دستاورد

۸	۲۸۸۶۲	۴۸۴	۱۴	دستاورد
۱۷۵	۴۸۷۰۸	۲۲۲۵	۴۵۰۲	آیین
۱۷	۴۸۷۰۸	۱۱۴	۱۷۵۵	آئین

همان‌طور که در نتایج جدول ۵ دیده می‌شود، تعداد کل نتایجی که پایگاه ایرانداک برای کاوش صورت‌های مختلف نوشتاری واژه‌های برگزیده بازیابی می‌کند؛ در مورد همه کلیدواژه‌ها (ویژگی‌های نوشتاری) متفاوت است. همچنین یافته‌های جدول ۵ در مورد پایگاه آی اس سی نیز حکایت از فقدان توجه کافی به ویژگی‌های نوشتاری فارسی دارد که اختلاف نتایج بازیابی شده گواه آن است. به‌عنوان نمونه موارد بارزی از این اختلاف را می‌توان در کلیدواژه‌هایی مانند "پتاسیم و پتاسیوم" مشاهده نمود.

از سوی دیگر نتایج حاصل از جستجو در پایگاه نورمگز نیز نشان می‌دهد تعداد نتایج بازیابی شده در مورد اغلب کلیدواژه‌ها (ویژگی‌های نوشتاری) متفاوت است؛ مانند آنچه در جستجوی کلیدواژه "فناوری و فن‌آوری" اتفاق می‌افتد که تعداد نتایج بازیابی شده برای این دو اختلاف زیادی با هم دارند. موارد استثناء نیز شامل کلیدواژه‌های "اساتید، استادها، استادان"، "زلزله بم، زلزله ی بم، زلزله بم" و "آیین، آئین" هستند که صورت‌های نوشتاری گوناگون در نتایج بازیابی شده دیده می‌شود.

در رابطه با پایگاه سید نیز اختلاف نتایج بازیابی شده برای صور گوناگون نگارشی واژه‌ها در بسیاری موارد باعث ایجاد اختلال و ارائه نتایج ناقص به کاربر می‌شود. به‌عنوان نمونه شاهد این اختلاف در مورد کلیدواژه‌های "پتاسیوم، پتاسیم" و "یا" دانشگاهها، دانشگاه‌ها" هستیم. همچنین سید در مورد کلیدواژه‌هایی مانند "مسکن، مُسکن و مَسکن" و "دستاورد و دستاورد" قادر به شناسایی اعراب‌گذاری و علائم بکار رفته نبوده و نتایجی با محتوای محل سکونت بازیابی می‌نماید. در دیگر پایگاه‌های بررسی شده نیز کاربر یا با اتفاقی مشابه آنچه گفته شد روبرو می‌شود و یا در صورت اعراب‌گذاری با نتایجی بدون ارتباط مواجه خواهد شد.

در مورد ویژگی‌های معنایی زبان فارسی نیز مشکلاتی وجود دارد که برخی مصادیق آن در ارتباط با پایگاه‌های اطلاعاتی مورد پژوهش در جدول شماره ۶ مشاهده می‌شود.

جدول ۶. نتایج بازیابی شده برای کلیدواژه های منتخب با معانی گوناگون^۱ در پایگاه های مورد مطالعه

مجموع نتایج بازیابی در پایگاه های اطلاعاتی				کلیدواژه مورد جستجو	
سید	نورمگز	آی اس سی	ایراندک		
۶	۷	۱۴	۱۰	توپ بازی	توپ
۴	۱۲	۳	۵	توپ جنگی	
۰	۰	۰	۰	واحد شمارش	
۶	۰	۳	۵	پارچه	
۰	۱	۰	۰	قسمتی از یک آنتی ژن (علم ژنتیک)	
۰	۱	۰	۰	معنای عوامانه	
۲	۱۴	۱	۱	شوق	شور
۱۶	۳	۱۷	۱۸	شور بودن	
۲	۳	۰	۰	مشورت	
۰	۰	۲	۱	غیر مرتبط	
۲۰	۲۰	۱۸	۱۶	روح	روان
۰	۰	۲	۴	جاری	
۶۵۱	۵۸۰۳۲	۱۶۵۴	۲۳۶۰	دریای خزر	
۳	۵۶۸۹۷	۲۵	۲۷	دریای کاسپین	

^۱در مورد ویژگی های معنایی بدلیل نبودن حالت های مختلف نوشتاری کلمات، (بجز کلید واژه های مربوط به دریای مازندران) معانی مختلف بازیابی شده در برای هر کلیدواژه، در بین ۲۰ نتیجه ی نخست بازیابی شده شمارش و ثبت شدند.

۵۴	۶۳۸۸۲	۱۷۰	۴۳۹	دریای مازندران
----	-------	-----	-----	----------------

همان‌طور که در جدول ۶ مشهود است، با توجه به نتایج حاصل از کاوش کلیدواژه‌ها در پایگاه ایرانداک شاهد عدم توجه پایگاه‌های اطلاعاتی به واژگانی که دارای معانی مختلف و صورت نوشتاری یکسانی هستیم که موجب می‌شود گاه معانی مختلفی از یک واژه بازیابی شود، اما معنی مورد نظر کاربر در میان آنها نباشد. همچنین نتایج نشان می‌دهد که پایگاه آی اس سی نیز توجه کافی به ویژگی‌های یاد شده جهت تأمین نیازهای کاربران فارسی‌زبان در رابطه با توجه به ویژگی‌های معنانشناسی زبان فارسی ندارد، به‌عنوان مثال در مورد کلیدواژه "دریای مازندران" اختلاف قابل توجهی میان تعداد نتایج و مدارک بازیابی شده برای کلیدواژه‌های مترادف آن وجود دارد که جامعیت کاوش را تحت تأثیر قرار می‌دهد. همچنین، مانند نتایج دیگر پایگاه‌ها در بخش کاوش معنایی، در مورد واژه‌هایی مانند "توپ" و "شور" کاربر با بازیابی نتایج نامرتب مواجه شده و یا ممکن است هرگز بدون استفاده از کلیدواژه کامل‌کننده به نتایج مورد نظرش دست پیدا نکند.

علاوه بر موارد یاد شده، با دقت در نتایج حاصل در می‌یابیم که در پایگاه اطلاعاتی نورمگز نیز مشکلاتی وجود دارد. چراکه گاهی کاربران در جستجوی معانی مختلف یک واژه دچار مشکل شده و از بازیابی نتایج مطلوب باز می‌مانند و یا با نتایج نامربوط مواجه می‌شوند. مثلاً بازیابی نتایجی با معنای "مشورت و هم‌اندیشی" برای جستجوی کلیدواژه "شور" که در دیگر پایگاه‌ها نتایجی با این مضمون بازیابی نشد که بسته به نیاز کاربر می‌تواند مطلوب یا نامطلوب باشد. البته ذکر این نکته ضروری است که در قسمت واژگان مترادف از بخش معنایی، این پایگاه ضمن بازیابی نتایجی نزدیک به هم برای کلیدواژه‌های "دریای خزر، دریای مازندران و دریای کاسپین"، پس از بازیابی نتایج کاوش برای هر یک از کلیدواژه‌های "دریای خزر و مازندران" دیگری را نیز برای کاوش به کاربر پیشنهاد می‌دهد که خود گامی مؤثر در بهبود بازیابی کلیدواژه‌های مترادف است.

همچنین بر اساس نتایج بازیابی شده، پایگاه سید نیز توجه کافی را به ویژگی‌های معنایی زبان فارسی نداشته است. چنانچه در مورد ویژگی ترادف در واژه‌هایی که به "دریای مازندران" اشاره دارند، اختلاف زیادی در تعداد نتایج بازیابی شده وجود دارد.

۵. بحث و نتیجه‌گیری

هر یک از ویژگی‌های نوشتاری و معنایی زبان فارسی که در این پژوهش ذکر شد به نحوی چالش‌هایی را فرا روی کاربران فارسی‌زبان در هنگام کاوش و بازیابی اطلاعات از منابع مختلف مانند موتورهای کاوش وب و پایگاه‌های اطلاعاتی قرار می‌دهد. علاوه بر نتایج و مصادیق ذکر شده که تأثیر قابل توجه خصوصیات یاد شده را نشان می‌دهند؛ شرحی از این مشکلات به همراه مصادیق آنها بر پایه یافته‌های پژوهش‌های انجام شده نیز، می‌تواند صحه‌ای بر نتایج بدست آمده در این پژوهش باشد. برای این منظور و بدست آوردن تصویر بهتری از مسائل و چالش‌های زبان فارسی در زمینه جستجو از پایگاه‌های اطلاعاتی ابتدا شرحی از ویژگی‌های نوشتاری و سپس ویژگی‌های معنایی بیان می‌شود.

ویژگی‌های نوشتاری

در بخش مسائل نوشتاری، نتایج پژوهش‌های آخشیک و فناعی (۱۳۹۱) و گل تاجی و بذرگر (۱۳۸۹) نشان داد که چالش‌های ریختی زبان فارسی از جمله پیوسته‌نویسی و جدانویسی واژگانی مانند واژگان مشتق و مشتق-مرکب و برخی ویژگی‌های مندرج در جدول ۳، تأثیر زیادی بر بازیابی اطلاعات از پایگاه‌های اطلاعاتی داخلی از جمله پایگاه آی اس سی، ایرانداک و سید دارد که نتایج جدول شماره ۵ نیز مؤید آن است. بدین ترتیب می‌توان به وضوح دریافت که حتی پایگاه‌های داخلی فارسی‌زبان نیز توجه کافی را به ملزومات و خصیصه‌های زبان فارسی ندارند. کاربران این قبیل پایگاه‌ها که در بسیاری موارد در پی جستجوی اطلاعات جهت انجام کارهای پژوهشی و مطالعاتی هستند به واسطه‌ی این مشکل از دستیابی جامع به اطلاعات مورد نیازشان باز می‌مانند. مانند آنچه در جستجوی واژه "فناوری" و "فن‌آوری" بدست آمد. این در حالی است که کاربران در سطوح مختلف و با عادت‌های نوشتاری گوناگون ممکن است هریک از این صورت‌ها را برای جستجو استفاده کنند و در بسیاری موارد ناخودآگاه از دستیابی به نتایجی که حاصل کاوش شکل دیگر واژه هستند محروم بمانند و به همان نتایج اولیه بسنده کنند. البته موارد استثنایی هم وجود دارد؛ چنانکه به نظر می‌رسد با توجه به بازیابی تعداد مساوی و یا نزدیک نتایج برای صورت‌های مختلف برخی کلمات و همچنین دیده‌شدن صورت‌های مختلف هر یک از واژه‌ها در پی کاوش تنها یک صورت از آن در بین نتایج بازیابی شده، پایگاه نورمگز در بعضی موارد فارغ از صور نوشتاری گوناگون و بر مبنای مفهوم واژه‌ها عمل نموده و همه منابع حاوی آنها را در یک دسته معنایی قرار داده است.

در مجموع شناسایی آن دسته از ویژگی‌های زبانی که نقش پررنگ‌تر و عمده‌ای در بروز چالش‌های پیش روی کاربران دارند نیز نکته‌ای است که نباید از آن غافل شد. چنانکه هم‌اوندی (۱۳۹۲) در پژوهش خود مواردی مانند پیوسته‌نویسی و جدانویسی واژگان مشتق، مشتق-مرکب،

انواع جمع‌های فارسی و مکسرِ عربی، نگارش همزه بدونِ کرسی و استفاده از زبان محاوره را از مشکلات عمده زبان فارسی در جستجو و بازیابی تصاویر خاصه در موتورهای کاوش می‌داند. همچنین یافته‌های پژوهش عبدالهی نورعلی (۱۳۸۸) نشان می‌دهد بین شکل نوشتاری واژه و ابزار جستجو (موتورهای کاوش) رابطه معناداری وجود دارد. بنابراین، می‌توان نتیجه گرفت که بکار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جستجوی خاص در بازیابی اطلاعات اثرگذار است. به‌عنوان نمونه چنانچه کاربری کلیدواژه "باغها" را با شکل پیوسته انتخاب کند، بیشتر اطلاعات موجود که با کلیدواژه "باغ‌ها" نمایه‌سازی و ذخیره شده‌اند را از دست می‌دهد و از طرف دیگر انتخاب کلمه "باغ‌ها" نیز منجر به بازیافت نتایج نامربوط می‌شود. همچنین است تفاوت بین تعداد نتایج بازیابی شده حاصل از کاوش کلیدواژه‌های "پرستش گاه و پرستشگاه"، "فناوری و فن‌آوری" و "بزرگسال و بزرگ سال" که در جدول شماره ۴ مشاهده شد. در این بین فقدان توجه و آگاهی کاربر از این مسئله می‌تواند باعث عدم دسترسی وی به نتایج حاوی صورت دوم واژه شود. همان‌طور که مشهود است وجود مسئله پیوسته‌نویسی و جدانویسی در نتایج همه‌ی پژوهش‌های یاد شده، گواهی بر تأثیر زیاد این خصیصه بر کاوش و بازیابی اطلاعات به زبان فارسی است. این در حالی است که پیوسته‌نویسی و جدانویسی واژگان در بسیاری موارد به‌صورت سلیقه‌ای انجام شده و خیلی از نویسندگان از قواعد مربوط به زبان در مورد آن پیروی نمی‌کنند. وجود مسائلی مانند استفاده از فاصله و نیم‌فاصله و جدا در نظر گرفته شدن اجزای یک واژه مرکب مانند "کتاب‌خانه" و بازیابی نتایج مستقل برای هر یک از این اجزاء نیز موجب تشدید این موارد می‌شود. در مجموع در بخش نوشتاری، با دقت در نتایج حاصل از پژوهش می‌توان دریافت که علاوه بر مشکلات مربوط به پیوسته‌نویسی و جدانویسی واژگان، کلمات عربی موجود در زبان فارسی مانند انواع جمع‌ها و انواع شیوه‌های نگارش همزه، واژگان دخیل و معادل آنها سهم زیادی از مشکلات پیش‌روی کاربران فارسی‌زبان را در تعامل با محیط‌های اطلاعاتی و خاصه پایگاه‌های اطلاعاتی فارسی، به خود اختصاص داده‌اند. به‌طور کلی اختلاف نتایج بازیابی شده برای حالات مختلف نوشتاری که در اکثر موارد در همه پایگاه‌ها مشابه است به نوعی حاکی از الگوی نوشتاری غالب در بین نویسندگان منابع هر پایگاه است و می‌توان از آن برای استخراج الگو به‌منظور نمایه‌سازی یا پیشنهاد کلیدواژه کاوش استفاده نمود؛ مثل واژه‌های مرکب "بزرگسال و بزرگ‌سال" که در هر چهار پایگاه مورد بررسی جستجوی صورت جدای آن منجر به بازیابی نتایج بیشتری از فرم پیوسته آن شد. همچنین در بسیاری موارد که صورت‌های نوشتاری مختلف واژه‌ها دارای گونه‌های با اعراب و بدون اعراب‌گذاری بوده‌اند، به نظر می‌رسد که ترجیح نویسندگان به استفاده از صورت‌های بدون اعراب است که می‌تواند ریشه در عادات نوشتاری فارسی‌زبانان و عدم تمایل آنها برای استفاده از

اعراب‌گذاری و همچنین پیچیده بودن اعراب‌گذاری در صفحه کلیدها اشاره نمود مانند نتایج بازیابی شده برای واژه‌های "معلم و مسکن".

ویژگی‌های معنایی

با دقت در پژوهش‌های یاد شده می‌توان دریافت که ابعاد معنایی زبان در آنها مهجور مانده، در حالی که ویژگی‌های خاص معنایی زبان فارسی نقش قابل توجهی در مسئله کاوش و بازیابی اطلاعات داشته و سبب ایجاد مشکلاتی در تعامل بین کاربران و ابزارهای اطلاع‌رسانی هستند. همچنان که نتایج پژوهش هم‌اوندی (۱۳۹۲) با نتایج حاضر همسو است. در این صورت جستجوی کاربر با مسائلی مانند بازیافت نتایج نامربوط مواجه می‌شود و در مواردی حصول نتیجه مطلوب مستلزم تکرار جستجو با شیوه‌های مختلف و واژگان تکمیلی است. مانند آنچه در مورد کاوش واژه "توپ" و یا واژه‌های "دریای خزر، دریای مازندران و دریای کاسپین" (در مورد اخیر پایگاه مجلات تخصصی نور نسبت به دیگر پایگاه‌ها، عملکرد قابل قبولی داشت) اتفاق افتاد (جدول شماره ۶). در مجموع در بخش معنایی برخی ویژگی‌ها مانند مترادف و همنامی و هم‌نویسگی، علیرغم کم توجهی به آنها در بسیاری از پژوهش‌های مربوطه، از جمله مشکلات جدی کاربران در استفاده از پایگاه‌های اطلاعاتی هستند. چه‌بسا اگر در فرایند نمایه‌سازی و کاوش، انتخاب کلیدواژه براساس یک اصطلاح‌نامه و با ارجاع به فرم صحیح و منتخب واژه انجام می‌شد تا حد زیادی به بهبود کاوش این دست واژه‌ها مؤثر بود.

آنچه ذکر شد خلاصه‌ای از پیامدهای کم‌توجهی و یا نادیده انگاشتن خصوصیات زبان فارسی توسط کاربران و طراحان موتورهای کاوش و پایگاه‌های اطلاعاتی است، که در نهایت منجر به اختلال در کاوش و بازیابی انواع گوناگون اطلاعات می‌شود. در مجموع می‌توان گفت که، پایگاه‌های اطلاعاتی (حتی انواع بومی آنها) نسبت به ویژگی‌های نوشتاری و معنایی زبان فارسی توجه کافی ندارند در حالی که بخش بزرگی از این مسائل را می‌توان به مرحله نمایه‌سازی مربوط دانست که باید طی آن ویژگی‌های زبان فارسی را مد نظر قرار دهند و با تمهیداتی از جمله یکدست‌سازی واژگان و استفاده از اصطلاح‌نامه‌ها یاد نسبت به حل آن اقدام نمایند. این مسئله موجب می‌شود که احتیاج فارسی‌زبانان بویژه به ابزارهای کاوش بومی و پایگاه‌های اطلاعاتی که مبتنی بر ویژگی‌های زبانی خودشان طراحی شده باشد و در تعامل با کاربران فارسی‌زبان به ظرائف و باید و نبایدهای زبانی آنها بیشتر توجه کند، بیش از پیش احساس شود. از سوی دیگر پایگاه‌های اطلاعاتی نیز باید نسبت به برآورده ساختن نیازهای زبانی کاربرانشان و اصلاح تعاملشان با آنان بیشتر تلاش کنند. چراکه، نادیده انگاشتن و یا کم‌توجهی به شاخصه‌ها و ویژگی‌های زبانی

کاربران موجب بروز مسائلی در امر جستجو و بازیابی اطلاعات می‌شود که در نهایت از دست رفتن اطلاعات مفید و یا بازیابی اطلاعات ناخواسته را به همراه خواهد داشت. با توجه به یافته‌های پژوهش، پیشنهادهای زیر می‌تواند در پیشگیری و رفع این چالش‌ها و بهبود تعامل میان کاربران و ابزارهای کاوش و بازیابی اطلاعات مؤثر باشد:

- بر اساس یافته‌های پژوهش و با توجه به تأثیر قابل توجه صورت‌ها مختلف نوشتاری واژگان بر جستجو و بازیابی نتایج در پایگاه‌های اطلاعاتی، لازم است از ابتدا شیوه‌های نگارشی در محیط دیجیتال و حداقل در مورد متون علمی و تخصصی تا حد امکان یکپارچه و هماهنگ باشند، وجود یک نرم‌افزار واژه‌پرداز منطبق با ویژگی‌های زبان فارسی^۱، می‌تواند یکی از راه‌های نیل به این مقصود باشد و در صورت بروز خطاهای نوشتاری و املائی به کاربر برای تصحیح آن هشدار دهد. طرح سامانه "استانداردساز و خطایاب متون فارسی" که توسط پژوهشکده مدیریت دانش ایرانداک به شورای پژوهش ارائه و تصویب شده و در دست اجرا است (که حاصل آن نرم‌افزاری برای خطایابی متون فارسی و برای ویرایش پایگاه گنج این پژوهشگاه خواهد بود) نمونه‌ای از این تلاش‌هاست.
- در مواردی مانند واژگان دخیل و معادل آنها که در پایگاه‌های مورد بررسی توجه لازم به آن صورت نگرفته بود، استفاده از واژگان معادل و تلاش برای رواج آن از طریق رسانه‌های مکتوب و غیرمکتوب، خصوصاً رسانه‌های رسمی از ابتدای ورود و استفاده از یک فناوری یا مفهوم (مثل واژه‌های یارانه و سوبسید^۲ و یا پیامک و اس ام اس^۳)، می‌تواند در پذیرش و کاربرد گسترده و هماهنگ آن توسط افراد مؤثر باشد.
- در مورد مسئله تنوع صورت‌های نوشتاری واژگان، پایگاه‌های اطلاعاتی می‌توانند از طریق فراهم ساختن نمایه‌های مناسب و استفاده از اصطلاح‌نامه‌ها و واژه‌نامه‌ها کاربران را از وجود صورت‌های مختلف نوشتاری یک واژه آگاه کنند و در صورت لزوم ارجاع بدهند. نمونه‌ای از این گونه تلاش‌ها را می‌توان در بخش تحقیق و توسعه ایرانداک مشاهده نمود. این پژوهشگاه حدود یک سال است که، نرم‌افزاری طراحی نموده که به وسیله آن نمایه‌سازان پژوهشگاه با دسترسی به مجموعه اصطلاح‌نامه‌های تدوین و ترجمه شده گروه اصطلاح‌شناسی این پژوهشگاه و بر اساس اصطلاح‌نامه‌ها و مجموعه واژگان‌های موجود نمایه‌سازی و سازمان‌دهی مدارک را انجام می‌دهند. این تلاش در

^۱-نرم افزار ویراستیار از جمله این نوع تلاش‌ها است.

^۲-Subsidy

^۳-Short Message Service

نهایت پشتمانه مناسبی برای کاوش کاربران و حل مسائل و چالش‌های

زبانی آنها خواهد بود.

- با توجه به یافته‌های پژوهش بخش عمده‌ای از مشکلات ناشی از ناهماهنگی در نوشتاری صورت‌های مختلف واژگان است، در نظر گرفتن ساز و کارهایی جهت اطلاع‌رسانی و آموزش به کاربران فارسی‌زبان چه در مقام نویسندگان در وب در جهت ایجاد یکدستی و هماهنگی در متون و چه در جایگاه کاوشگران اطلاعات در جهت آموزش ویژگی‌های خاص زبانی و انواع روش‌های بهبود کاوش می‌تواند در جلوگیری از بروز مشکلات یاد شده در پژوهش مؤثر باشد.
- جهت پوشش صورت‌های گوناگون نوشتاری و معنایی واژه‌ها، کاربرد نمایه‌سازی مشارکتی^۱ و توجه به آنچه کاربران مختلف از طیف‌های گوناگون و با عادات نوشتاری متفاوت پیشنهاد و جستجو می‌کنند، می‌تواند راهکاری جهت بهبود جستجو و بازیابی باشد.
- هوشمندسازی نمایه‌سازی و واژگان در پایگاه‌های اطلاعاتی در مرحله ذخیره و بازیابی و نیز هوشمندسازی الگوریتم‌ها و مدل‌های بازیابی اطلاعات (مانند آنچه گوگل در زمینه هماندها، خطاهای املائی، و ... انجام می‌دهد).

^۱-Folksonomy

فهرست منابع

- اسلامی، محرم ۱۳۸۱. دشواری‌های پردازش رایانه‌ای خط فارسی. فصلنامه نشر دانش ۱۹ (۳). <http://www.noormags.com/view/fa/articleage/47746> (دسترسی در ۱۳/۲/۱۳۹۳).
- اکبری نژاد، سعید ۱۳۷۶. فاصله خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات. فصلنامه کتاب ۸ (۱ و ۲). <http://www.noormags.com/view/fa/articlepage/87024> (دسترسی در ۱۳/۲/۱۳۹۳).
- آخشیک، سمیه سادات و رحمت الله فتاحی. ۱۳۹۱. تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی. فصلنامه کتابداری و اطلاع رسانی ۱۶ (۳): ۹-۳۰.
- حری، عباس. ۱۳۷۲. کامپیوتر و رسم الخط فارسی. فصلنامه تحقیقات اطلاع رسانی و کتابخانه‌های عمومی ۱۳ (۱). <http://www.noormags.com/view/fa/articlepage/396231> (دسترسی در ۱۳/۲/۱۳۹۳).
- حسینی‌بهشتی، ملوک‌السادات ۱۳۸۶. معنی‌شناسی واژگانی فرااصطلاحنامه و بازیابی اطلاعات. کتاب ماه کلیات مجموعه اطلاع رسانی و کتابداری ۱۰ (۱۰): ۳۰-۳۷.
- رائی ساربانقلی، محمد. ۱۳۸۵. مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی: کاربران مرکز اینترنت دانشگاه اسلامی واحد شبستر. فصلنامه کتاب ۱۷ (۳). <http://www.noormags.com/view/fa/articlepage/159553> (دسترسی در ۱/۶/۹۱).
- رسولی، محمد صادق و بهروز مینایی بیدگلی. ۱۳۸۷. روشی جدید در خطایابی املایی در زبان فارسی. دومین کنفرانس داده کاوی ایران، تهران، ۲۲-۲۱ آبان ۱۳۸۷.
- روح پرور، رحیمه و محمود بی جن خان. ۱۳۸۶. به کارگیری یک نظام برجسب دهی برای تعبیر و تفسیر یک پیکره متنی زبان فارسی. هفتمین همایش زبان‌شناسی ایران، تهران، آذر ۲۰-۲۱، ۱۳۸۶.
- زره ساز، محمد و رحمت الله فتاحی. ۱۳۸۵. ملاحظات اساسی در طراحی رابط کاربر نظام‌های رایانه‌ای و پایگاه‌های اطلاعاتی. مطالعات ملی کتابداری و سازماندهی اطلاعات. دوره ۱۷. شماره ۲. صفحه ۲۵۱-۲۶۸.
- ستوده، هاجر و زهره هنرجویان. ۱۳۹۱. مروری بر دشواریهای زبان فارسی در محیط دیجیتال و تاثیرات آنها بر اثر بخشی پردازش خودکار متن و بازیابی اطلاعات. فصلنامه کتابداری و اطلاع رسانی. دوره ۱۵. شماره ۴. صفحه ۵۹-۹۲.
- عبداللهی نورعلی، محمد صادق و عبدالرسول جوکار. ۱۳۸۸. چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب. مطالعات تربیتی و روانشناسی. ۱۰ (۲): ۶۷-۹۰.
- گل تاجی، مرضیه و سعیده بذرگر. ۱۳۸۹. بررسی مشکلات ریخت شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی. فصلنامه کتابداری و اطلاع رسانی. ۱۳ (۲): ۱۹۹-۲۲۲.

لیلا مرتضائی . مسائل زبان و خط فارسی در ذخیره سازی و بازیابی اطلاعات. پژوهشنامه

پژوهش و مدیریت اطلاعات. ۱۳۸۰؛ ۱۷ (۱ و ۲): ۱۹-۲۶

نشاط، نرگس. ۱۳۷۹. مسائل رسم الخط فارسی در رویارویی با فناوری نوین اطلاعاتی. در فهرست های رایانه ای،

کاربرد و توسعه. مجموعه مقالات همایش کاربرد و توسعه فهرست های رایانه ای در کتابخانه های ایران،

آبان ۲۷-۲۸، (۴۰۱-۴۰۸). مشهد: دانشگاه فردوسی مشهد.

هماوندی، هدی. ۱۳۹۲. بررسی مشکلات جستجوی بازیابی تصاویر در موتورهای کاوش برگزیده مبتنی بر ویژگی های

نگارشی زبان فارسی، پایان نامه کارشناسی ارشد، دانشگاه قم.

Hammo, B. H. 2009. Towards enhancing retrieval effectiveness of search engines for diacritized Arabic documents. *Information Retrieval* 12 (3). <http://link.springer.com/article/10.1007/s10791-008-9081-9>. (Accessed July 19, 2012).

Internet world stats: Usage and population statistics, 2015 November 30. <http://www.internetworldstats.com/stats7.htm>. (Accessed January 20, 2016).

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). <http://www.emeraldinsight.com/journals.htm?articleid=1640585>. (Accessed July 17, 2012).

Lazarinis, F. 2008. Improving concept-based web image retrieval by mixing semantically similar Greek queries. *Program: electronic library and information systems* 42 (1). <http://www.emeraldinsight.com/journals.htm?articleid=1674242>. (Accessed July 17, 2012).

Lewandowski, D. 2008. Problems with the use of Web search engines to find results in foreign languages. *Online Information Review* 32(4). <http://www.emeraldinsight.com/journals.htm?articleid=1747662>. (Accessed June 15, 2012).

Marchionini, G. 2008. Human information interaction research and development. *Library and information science research* 30 (4). http://ils.unc.edu/~march/Marchionini_Inf_interact_LISR_2008.pdf. (Accessed May 4, 2014)

Ruiter, De. J. 2006. Natural Language Interaction - the understanding computer. Essay for the course: *Human Computer Interaction*. http://www.jdruiter.nl/published_work/Natural_language_interaction.pdf. (Accessed June 10, 2014)

Zhang, J and Suyu L. 2007. Multiple language supports in search engines. *Online Information Review* 31 (4). <http://www.emeraldinsight.com/journals.htm?articleid=1621798>. (Accessed July 13, 2012).

Survey of Information Searching and Retrieving Challenges in Databases in Connection with Persian Language Writing Features

Hoda Homavandi¹ | Yaghub Norouzi² |
Moluk S. Hoseine Beheshti³

1. PhD Candidate in Knowledge & Information Science, Tehran University
h.homavandi@ut.ac.ir

2. PhD in Knowledge & Information Science, Associate Professor
;Qom University



ynorouzi@gmail.com

3. PhD in General Linguistics, Assistant Professor; Iranian Research Institute for Information Science & Technology (IRANDOC), Tehran, Iran.

beheshti@irandoc.ac.ir

Abstract

Purpose: The present research was carried out with the aim of explicating the major writing and semantic problems of Persian language when using data environments and determining the degree of compatibility and attention to these features in Persian databases.

Methodology/Approach: The present research is of survey analytical type being conducted through direct observation. Having reviewed the related literature, we kept a checklist of search key words. Each of these key words was searched in the databases under study, such as Iranian Research Institute for Information Science and Technology, regional Centre for Information Science and Technology, Noor Magaz, and Scientific Information database Affiliated with Jihad Daneshgahi, and the number of retrieved findings was recorded.

Findings: Some of the writing and semantic features of Persian language contribute to problems associated with retrieving information from the selected databases. Some of these features include connected and disconnected forms of writing of derivative, compound, and derivative-compound words, diversity of plural forms, loanwords and their equivalents in writing as well as polysemy, homonymy, etc., in semantics. For instance, retrieving different results for various writing forms of the key words "فناوری و فن آوری" as derivative-compound words or "پتاسیوم و پتاسیم" as various forms of recording words, or retrieving different findings for key words "دریای خزر، دریای مازندران و دریای کاسپین" as well as lack of their appropriate coverage as synonymous words and giving the user information about it in order to improve the exploration process, for it has negative effects on search and retrieval process.

Conclusion: Findings indicated that Persian databases do not pay adequate attention to writing and semantic features of Persian language, and disregard many of its features in searching and retrieving information. In connection with the impact of these features on the interaction of users with databases, Persian-speaking users' need for native exploration tools and databases designed in

accordance with the features of their own language have become more and more urgent. The present research has examined the ability of Persian databases in covering some of the features of this language, which have a noticeable impact on the process of searching and retrieval, pinpointing the weak points and strengths of these databases. The results of the present research could be utilized to improve the performance of the above-mentioned databases.

Keywords: information retrieval; Databases; Persian language; Writing features