اولین همایش ملی جویشگر بومی
۲۷ و ۲۸ بهمن ماه سال ۱۳۹٤
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
وزارت ارتباطات و فناوری اطلاعات
ITRC

صهبا: جستجوگر اشعار فارسی

سید مرتضی خالقی[۱]، راضیه فرجام‌فرد[۲]

[۱] کارشناسی‌ارشد مهندسی فناوری اطلاعات، شرکت کارانس ایرانیان
morteza.khaleghi@karans.co
[۲]کارشناسی‌ارشد مدیریت کسب و کار، شرکت کارانس ایرانیان

چکیده

امروزه وب به عنوان منبعی عظیم، اطلاعاتی از حوزه‌های گوناگون را در خود جای داده است. اگرچه به نظر می‌رسد کاربران با استفاده از جستجوگرهای موجود، مشکلی در یافتن اطلاعات نداشته باشند، اما تنوع حوزه‌ها سبب می‌شود یافتن اطلاعات در حوزه‌های تخصصی به امری دشوار تبدیل شود. بنابراین برای یافتن اطلاعات در حوزه‌های تخصصی به جستجوگرهای مختص به حوزه نیاز است. با وجود نمونه‌های موفق جستجوگرهای موضوعی در داخل و خارج از کشور، حوزه اشعار فارسی مورد توجه قرار نگرفته است. این پژوهش ابتدا به بررسی چالش‌های شعر فارسی پرداخته و در ادامه با ارائه معماری پیشنهادی به معرفی جستجوگر شعر فارسی (صهبا) پرداخته است.

کلمات کلیدی

موتور جستجو، جویشگر، شعر فارسی، خزنده وب

۱- مقدمه

امروزه وب به منبعی غنی از اطلاعات تبدیل شده که حوزه های گوناگونی از جمله موسیقی، تاریخ، ورزش، علم، بهداشت، ادبیات و.. را در خود جای داده است. این موضوع سبب شده تا یافتن اطلاعات مطلوب برای کاربران به امری دشوار تبدیل شود. زیرا هر جستجو در جستجوگرهای همه منظوره مانند گوگل (www.google.com) و بینگ (www.bing.com) چندین هزار نتیجه خواهد داشت [1].

در جستجوگرهای همه منظوره کاربران نمی‌توانند به صراحت حوزه (مثلا بهداشت یا ادبیات) مورد نظر خود را مشخص کنند بنابراین ممکن است در نتایج جستجو مطالبی خارج از حوزه مورد انتظار کاربر نیز وجود داشته باشد. به عنوان مثال کاربر با جستجوی کلمه "پیتزا" صفحاتی در مورد پیتزا و نحوه پخت آن را دریافت می‌کند اما احتمال دارد برخی از نتایج مربوط به رستوران‌های فروشنده پیتزا باشد که این نتایج مورد انتظار کاربر نبوده است. بنابراین درنهایت خود کاربر باید نتایج را مرور کند و نتایج مرتبط به جستجوی خود را بیابد که این کار نیازمند صرف دقت زیادی توسط کاربر است. در مواردی که کاربر حوزه نتایج مورد نظر خود را می‌داند جستجوگرهای که روی حوزه خاصی متمرکز هستند، می‌توانند کارگشا باشند.

جستجوگر مختص به حوزه[۱] که به آن جستجوگر موضوعی[۲] و یا جستجوگر عمودی[۳] نیز می گویند، برخلاف جستجوگرهای همه منظوره روی حوزه یا نوع خاصی از داده‌ها متمرکز هستند. محل تمرکز این جستجوگرها می تواند بر اساس حوزه داده‌ها (بهداشت، علم و..) و یا نوع داده (متن، فیلم، عکس و..) باشد. بنابراین جستجوگرهای موضوعی نیازی ندارند تا تمامی اطلاعات موجود در وب را جمع‌آوری و نگه‌داری کنند. کافی است فقط اطلاعات مرتبط با حوزه یا نوع داده‌ای مد نظر خود را دریافت کنند [2].

جستجوگرهای موضوعی در مقابل جستجوگرهای عمومی مزیت‌های زیادی دارند [3]. از جمله:

- دقت بالاتر در نتایج به دلیل محدود بودن حوزه
- امکان استفاده از دانش مختص به حوزه (مانند شبکه لغات[۴] و طبقه‌بندی معنایی[۵])
- امکان پالایش نتایج بر اساس پارامترهای مختص به حوزه
- امکان برقراری روابط (بالا به پایین، شبکه‌ای و..) بین موجودیت‌های حاضر در نتایج جستجو
- امکان اصلاح و بهبود پرس‌وجوی[۶] کاربر با توجه به حوزه
- هزینه پایین‌تر برای جمع‌آوری و نگه‌داری اطلاعات

امروزه جستجوگرهای همه منظوره مانند گوگل و بینگ کاربران زیادی را به خود جذب کرده‌اند، اما با توجه به مزایای ذکر شده، جستجوگرهای موضوعی گوناگونی وجود دارند که هر یک با تمرکز بر حوزه‌ای خاص به نیاز مخاطبان خود پاسخ می‌دهند. برخی از نمونه‌های موفق این جستجوگرها در جدول ۱ آمده است.

جدول (۱) : نمونه‌های جستجوگر موضوعی

آدرس اینترنتی	حوزه
yummly.com	آشپزی
quetzal-search.info	پزشکی
scholar.google.com	منابع علمی

علاوه بر نمونه‌های ذکر شده در کشور ایران نیز جستجوگرهای موضوعی برای برخی حوزه‌ها ایجاد شده‌اند که تعدادی از آن‌ها در جدول ۲ آمده است.

جدول (۲) : نمونه‌های داخلی جستجوگر موضوعی

آدرس اینترنتی	حوزه
khabarfarsi.com	اخبار فارسی
salamatjoo.ir	سلامت و تندرستی
parsquran.com	قرآن

در این میان به نظر می‌رسد حوزه زبان و ادبیات فارسی به خصوص گنجینه اشعار فارسی مهجور مانده است. با وجود منابع گوناگون از اشعار فارسی در وب، نیاز به یک مرجع واحد برای دسترسی به انواع شعر فارسی احساس می‌شود. در این پژوهش طرحی برای یک جستجوگر شعر و چالش‌های پیش روی آن را معرفی می‌کنیم.

در ادامه مقاله ابتدا به بررسی برخی از کارهای انجام شده در زمینه جستجوی اشعار می-پردازیم، پس از آن در بخش ۳ چالش‌های زبان فارسی برای جستجو به ویژه چالش‌های جستجوی شعر را بیان می‌کنیم و بخش ۴ به معرفی ساختار پیشنهادی برای جستجوگر اشعار فارسی اختصاص دارد.

۲- کارهای پیشین

تحقیقات نشان می‌دهد نیاز اطلاعاتی افراد در هنگام جستجو به سه دسته کلی تقسیم می‌شود[4]:

۱. پیمایشی[۷]: کاربر می‌داند به دنبال چه چیزی می‌گردد و فقط برای یافتن آدرس آن به جستجوگر مراجعه می‌کند.
۲. کسب اطلاعات[۸]: کاربر به دنبال مطلبی با موضوع خاصی می‌گردد.
۳. یافتن منابع[۹]: کاربر به دنبال منبع خاصی است تا آن را دریافت کند. مثلا به دنبال نرم افزاری برای دانلود است.

از میان نیازهای اطلاعاتی مطرح شده دو مورد اول در حوزه شعر متداول تر هستند. بنابراین کاربران به دو دلیل ممکن است نیاز به جستجو در اشعار داشته باشند:

۱. زمانی که قسمتی از یک شعر را در خاطر دارند و می خواهند به متن کامل آن دسترسی پیدا کنند.
۲. زمانی که نیاز به یافتن شعری با مضمونی خاص دارند.

تاکنون وب‌سایت‌های گوناگونی با هدف جمع‌آوری اشعار فارسی ساخته‌شده‌اند که می‌توان آن‌ها را به دو دسته کلی تقسیم کرد:

۱. سایت‌هایی که روی جمع‌آوری اشعار شاعران کهن فارسی تمرکز دارند (مانند سایت گنجور ganjoor.net).
۲. سایت‌هایی که شاعران معاصر می‌توانند با ساخت صفحه شخصی، اشعار خود را منتشر کنند (مانند irafta.com).

اگرچه نمونه‌های موفقی در هرکدام از دسته‌های ذکر شده وجود دارند اما رویکرد هیچ‌یک پاسخگویی به نیازهای اطلاعاتی عامه کاربران در زمان جستجو برای شعر، نبوده است. اگرچه تقریبا تمامی این وب‌سایت‌ها امکان جستجو را برای کاربران فراهم کرده‌اند؛ اما دقت نتایج جستجو راضی‌کننده نیست. مهم‌ترین دلیل این امر عدم اعمال پیش‌پردازش‌های لازم (یکسان سازی[۱۰]، ریشه‌یابی[۱۱] و حذف لغات پرکاربرد[۱۲]) است. از طرفی با توجه به تعداد زیاد این سایت‌ها جستجو در تک‌تک آن‌ها برای کاربران مطلوب نخواهد بود. بنابراین نیاز به یک جستجوگر موضوعی با تمرکز بر حوزه شعر احساس می‌شود. این جستجوگر علاوه بر جمع‌آوری اشعار از وب‌سایت‌ها، وبلاگ‌ها و سایر منابع، می بایست بتواند پیش‌پردازش‌های مورد نیاز برای شاخص‌گذاری[۱۳] و جستجو در اشعار فارسی را انجام دهد. با توجه به تفاوت‌هایی که در نگارش اشعار با متون معمولی وجود دارد و همچنین استفاده از لغات و اصطلاحات مرسوم در فارسی کهن، پیش‌پردازش اشعار نیازمند حل چالش‌های ویژه‌ای است.

۳- چالش‌های پیش‌رو

چالش‌های مربوط به موتور جستجوی شعر فارسی به دو بخش عمده تقسیم می‌شود:

۳-۱- چالش‌های زبان فارسی:

هر موتور جستجو باید قواعد مربوط به زبان جستجوی مورد استفاده در خود را رعایت کند. در موتورهای جستجوی زبان فارسی (خاص منظوره یا عام منظوره) نیز باید قواعد زبان فارسی رعایت شود. در زیر به بیان برخی از چالش‌های مربوط به زبان فارسی پرداخته شده است:

۳-۱-۱- ریشه‌یابی:

نظام های بازیابی اطلاعات در بیشتر زبان‌ها از امکانات ریشه یابی استفاده زیادی می کنند. به این معنا که با وارد کردن یک واژه به عنوان کلید واژه، به طور خودکار تمامی مشتقات واژه نیز جستجو می شود[5]. در برخی از زبان‌ها استفاده از الگوریتم هایی همچون پورتر[۱۴] و لاوینز[۱۵] برای ریشه یابی منجر به بهبود بخشیدن میزان دقت[۱۶] و فراخوانی[۱۷] شده است[6]. در زبان فارسی برای ریشه یابی افعال و کلمات ابتدا باید با چالش های پیش رو آشنا شد. چالش‌های مربوط به ریشه یابی در زبان فارسی به دو بخش عمده ریشه یابی اسم و صفت تقسیم می شوند:

۱.ریشه یابی اسم: در زبان فارسی، با حذف پسوندهای موجود در یک اسم می توان به ریشه آن دست پیدا کرد. در بیشتر پژوهش ها فرض شده است که پسوند اسمهای فارسی حداقل از دو حرف تشکیل شده است[6]. پسوندهای رایج و پرکاربرد در جدول ۳ آمده است. برای بدست آوردن ریشه اکثر کلماتی که با پسوندهای زیر جمع بسته می شوند؛ کافی است تنها پسوند آنها را حذف کرد. به عنوان مثال ریشه کلمه کتابم، کتاب و ریشه کلمه درختان، درخت می باشد. اما برای بدست آوردن ریشه برخی از این کلمات، صرفا حذف پسوند از انتهای آنها کافی نمی‌باشد؛ به عنوان مثال برای بدست آوردن مصدر کلماتی مانند کارخانجات و شبهات علاوه بر حذف پسوند "جات" و "ات"، اضافه کردن "ه" به اخر کلمه نیز نیاز است. به صورتی مشابه، کلمات مختوم به حروف صدادار نیز هنگام جمع بستن با ضمایر پیوسته دچار تغییراتی می‌شوند؛ به عنوان مثال برای جمع بستن کلمه "مو" با ضمیر "ـَم" نیاز به اضافه کردن حرف "ی" است: "مویم". بنابراین صرفا حذف کردن پسوند از انتهای این کلمات منجر به پیدا کردن مفرد آن‌ها نمی‌شود.

جدول ۳ : پسوند اسم‌ها

نوع پسوند	پسوند
ضمایر پیوسته	ـَم، ـَت، ـَش، ـِمان، ـِتان، ـِشان
جمع ساز فارسی	ها، ان، جات
جمع ساز عربی	ات، ین، ون

یکی دیگر از چالش های موجود در مورد پسوندهای نام برده شده، تشخیص ضمایر پیوسته در انتهای کلمات است. در انتهای برخی کلمات ممکن است حروف گفته شده در ضمایر پیوسته آمده باشد؛ در حالی که نقش این حروف در این کلمات ضمیر پیوسته نیست. به عنوان مثال حروف ـَم در انتهای کلمه "کتابم" نقش ضمیر پیوسته دارد و مصدر آن کتاب است. اما در کلمه ای مانند "رستم"، حروف □م قسمتی از خود کلمه است.

وجود جمع‌هایی مانند جمع‌های مکسر نیز یکی دیگر از چالش‌های مربوط به ریشه‌یابی است. جمع‌های مکسر کلماتی هستند که پیداکردن مفرد آن‌ها از هیچ قاعده خاصی پیروی نمی‌کند؛ به عنوان مثال مفرد کلمه "کتب" همان "کتاب" است.

۲.ریشه یابی صفت: در زبان فارسی برخی از صفت ها از ترکیب پسوند با اسم‌ها یا بن افعال بدست می آیند. پسوندهای رایج برای ساختن صفت در جدول ۴ آمده است. همانطور که دیده می‌شود بیشتر صفت‌های گفته شده از ترکیب بن مضارع با پسوند بدست می‌آیند. قاعده کلی برای بدست آوردن بن مضارع افعال، بدست آوردن حالت امری فعل برای سوم شخص مفرد و حذف "ب" از اول آن است. بدست آوردن حالت امری افعال از قاعده خاصی پیروی نمی‌کند. به عنوان مثال حالت امری فعل "آوردن"، با حذف "دن" بدست می‌آید در حالی‌که حالت امری فعل "رفتن"("رو")، از هیچ قاعده‌ای پیروی نمی‌کند. بنابراین برای پیدا کردن ریشه صفتهایی مانند "بیننده"، "بینا"، "گریان"، "آموزگار" باید راهی تعبیه کرد.

جدول ۴ : انواع صفت‌ها

نوع صفت	نحوه ساختن صفت
صفت فاعلی	۱. بن مضارع + ـَنده ۲. بن مضارع + ان ۳. بن مضارع + ا ۴. بن ماضی + ار ۵. بن فعل (یا اسم)+گار ۶. بن فعل (یا اسم)+گر ۷. بن فعل(یا اسم)+کار
صفت تفضیلی	صفت+تر
صفت برترین	صفت+ترین

۳-۱-۲- غلط املایی:

خطا در نوشتن کلمات موضوعی اجتناب ناپذیر است؛ به همین علت محققان زیادی تاکنون به بررسی غلطهای املایی در زبان‌های مختلف پرداخته‌اند[7][8][9]. این اشتباهات به دو دسته عمده تقسیم می‌شوند:

۱.اشتباهات املایی: اشتباهات املایی در اسناد دیجیتال ممکن است به علت نداشتن اطلاع در مورد شکل صحیح کلمات باشد. در زبان فارسی برخی حروف با یکدیگر "هم‌آوا[۱۸]" هستند. حروف هم‌آوا مانند هم تلفظ می‌شوند اما با یکدیگر متفاوتند؛ مانند حرف "س" و "ث" در کلمات "ساده" و "صابون". حروف هم‌آوا در جدول ۵ آمده است.

جدول ۵ : حروف هم‌آوا در زبان فارسی

گروه	اعضا
الف	آ، ا
ت	ط، ت
همزه	ی، ئ، و، ؤ، ا، أ، إ، ء
سین	س، ص، ث
زَین	ز، ض، ظ، ذ
ها	ح، ه
قاف	ق، غ

کلمات هم‌آوا نیز کلماتی هستند که تلفظ یکسان اما معانی متفاوت دارند؛ مانند کلمه "خویش"(به معنای خود، فامیل) و کلمه "خیش"(به معنای گاو آهن). برای پیدا کردن شکل صحیح این کلمات می‌توان از بستر معانی کلمات مجاور استفاده کرد. برخی حروف نیز"هم‌ریخت[۱۹]" می‌باشند. حروف یک ریخت حروفی هستند که احتمال اشتباه شدن آن‌ها هنگام نوشتن زیاد است؛ مانند حروف "ز" و "ژ". هنگام تایپ کردن حرف "ژ" اگر به درستی دکمه شیفت فشار داده نشود با حرف "ز" اشتباه خواهد شد[10]. حروف هم‌ریخت در جدول ۶ آمده است.

جدول ۶ : حروف هم‌ریخت در زبان فارسی

اعضا	گونه‌های هم‌ریختی
ا، أ، إ	تمامی گونه‌ها (اول، وسط و آخر)
پ، ب	تمامی گونه‌ها (اول، وسط و آخر)
ث، ت	تمامی گونه‌ها (اول، وسط و آخر)
ح، ج، چ	تمامی گونه‌ها (اول، وسط و آخر)
خ، ح	تمامی گونه‌ها (اول، وسط و آخر)
ذ، د	تمامی گونه‌ها (اول، وسط و آخر)
ژ، ز، ر	تمامی گونه‌ها (اول، وسط و آخر)
ض، ص	تمامی گونه‌ها (اول، وسط و آخر)
ظ، ط	تمامی گونه‌ها (اول، وسط و آخر)
غ، ع	تمامی گونه‌ها (اول، وسط و آخر)
گ، ک	تمامی گونه‌ها (اول، وسط و آخر)
ف، ق، ڤ	اول و وسط

۲.تک خطا[۲۰]: بیش از ۸۰٪ اشتباهات املایی به علت جایگزینی، حذف، اضافه و یا جابجایی اشتباه یک حرف در کلمه می‌باشد. این نوع از اشتباهات املایی را "تک خطا" گویند. ترکیب چند نوع از این خطاها در هنگام نوشتن کلمات را "خطای چندگانه[۲۱]" گویند[10]. نمونه تک خطا برای کلمه صلح در جدول ۷ آمده است.

جدول ۷ : انواع حالت‌های تک خطا (کلمه صحیح: صلح)

انواع تک خطا	مثال
جایگزینی	صاح
حذف	صح
اضافه	صلخح
جابجایی	صحل

۳-۲- چالش‌های شعر فارسی:

با توجه به خاص منظوره بودن موتور جستجوی پیشنهادی به حوزه شعر فارسی، چالش‌های مربوط به این حوزه نیز باید مورد توجه قرار گیرد. این چالش‌ها به صورت زیر است:

- پیدا کردن شعری با مضمون خاص یکی از نیازهای گفته شده در جستجوی شعر است. برای بازیابی اشعار مرتبط با یک مضمون خاص باید سعی شود کلمات موجود در عبارت جستجو در کمترین فاصله نسبت به هم ظاهر شوند. در نظرسنجی که از متخصصان حوزه ادب و شعر انجام شد؛ اشعاری که کلمات مورد نظر در آن‌ها با فاصله بیشتری قرار دارند؛ کمتر حاوی مضمون مورد نظر هستند.
- گاهی برای هماهنگ شدن کلمات در شعر به گونه‌ای که شاعر می‌خواهد؛ شکل ظاهری آن‌ها تغییر می‌کند؛ به عنوان مثال کلمه "ترا" که مخفف شده کلمه "تو را" است . این تغییر ظاهری باعث می‌شود برخی اشعار مرتبط در نتیجه جستجو ظاهر نشوند. گاهی نیز تغییر ظاهری کلمات منجر به تغییر معنی آن‌ها نیز می‌شود؛ به عنوان مثال کلمه "ماست" که مخفف شده کلمه "ما است" می‌باشد.
- در مجموعه اشعار موجود برخی کلمات دارای تعداد دفعات تکرار زیاد می‌باشند؛ مانند "از"، "به"، "و"، "را". این کلمات ارزش کمی در بازیابی اسناد مرتبط با جستجوی کاربر دارند. شناسایی لیستی از این کلمات و دخیل کردن آن‌ها در هنگام جستجو منجر به بهبود بخشیدن نتایج بازیابی اطلاعات خواهد شد.

۴- سامانه پیشنهادی

صهبا[۲۲] جستجوگر اشعار فارسی است که امکان جستجوی دقیق در مجموعه اشعار کهن و معاصر را در اختیار کاربران قرار می‌دهد. حتی اگر آشنایی کاربر با زبان فارسی به ویژه ادبیات کهن کافی نباشد، صهبا تلاش می‌کند با بهبود و اصلاح پرس‌وجوی کاربر همواره نتایج مطلوبی برای او فراهم کند.

۴-۱- معماری جستجوگر

معماری پیشنهادی برای جستجوگر صهبا از ۷ مولفه تشکیل شده است.

۴-۱-۱- خزنده

خزنده وظیفه جمع‌آوری اطلاعات از منابع مختلف را به عهده دارد، این منابع می توانند شامل وب‌سایت‌ها، وبلاگ‌ها، شبکه‌های اجتماعی و حتی پیام‌رسان‌های تلفن‌های همراه باشند.

۴-۱-۲- پیش‌پردازش اشعار

اشعار پیش از ورود به مخزن می بایست پاکسازی و یکسان سازی شوند تا در هنگام جستجو کاربران بتوانند نتایج مطلوبی دریافت کنند. این بخش از جستجوگر صهبا اشعار را از خزنده دریافت می‌کند و پس از انجام پیش‌پردازش‌های مورد نیاز آن‌ها را برای شاخص‌گذاری وارد مخزن می‌کند. پیش‌پردازش اشعار شامل دو گام اصلی می‌شود:

۱.یکسان‌سازی حروف: با توجه به اینکه برخی از حروف مانند "ک " و "ی" دارای کدهای کاراکتری متفاوتی هستند، برای تطبیق صحیح لغات می‌بایست آن‌ها را یکسان‌سازی کنیم.

۲.ریشه‌یابی: هر لغت می‌تواند به صورت‌های مختلفی در متن ظاهر شود. ازآنجایی‌که همه این صورت‌های ظاهری به مفهوم مشترکی که ریشه اسم یا فعل است، اشاره می‌کنند برای تطبیق صحیح ارتباط بین لغات باید صورت‌های مختلف آن‌ها را به ریشه برگردانیم. به‌عنوان‌مثال کلمات "کتاب‌ها"، "کتابم"،"کتابی" و.. همگی باید به ریشه "کتاب" برگردانده شوند. همچنین افعالی مانند "می‌دانند"،"دانستم"،"بدانیم"و.. نیز باید به ریشه "دانستن" برگردانده شوند.

در [6] برای بدست آوردن ریشه اسم‌ها از طراحی یک ماشین تعیین‌پذیر حالات متناهی[۲۳] (DFA) استفاده شده است. راه پیشنهادی در این پژوهش قادر به پیدا کردن ریشه جمع‌های مکسر و دیگر حالات خاص که در بخش قبل ذکر شد؛ نمی‌باشد. در [7] برای رفع مشکلات پژوهشهای قبل، از یک الگوریتم سلسله مراتبی استفاده شده است. در این الگوریتم ابتدا با استفاده از یک برچسب گذار واژگانی کلام[۲۴] (POS Tagger) برای تشخیص فعل، اسم و صفت استفاده شده است. سپس هر یک از اجزا برای ریشه یابی به یک DFA وارد می‌شوند. مزیت این الگوریتم در ریشه‌یابی حالات خاص مانند جمع‌های مکسر است. اگرچه استفاده از POS Tagger به عنوان پیش پردازش می‌تواند مفید باشد؛ اما ممکن است سرعت بازیابی اطلاعات را کم کند. ریشه یاب استفاده شده در جستجوگر صهبا بدون استفاده از POS Tagger می‌تواند فعل، اسم و صفت را تشخیص دهد. برای ریشه یابی اسم و صفت از یک DFA استفاده شده است. مشکلات گفته شده برای حالات خاص کلمات که در بخش "ریشه‌یابی اسم" و "ریشه‌یابی صفت" به آن اشاره شد؛ با استفاده از راه پیشنهادی قابل حل است.

۴-۱-۳- مخزن

مخزن هسته اصلی سامانه خواهد بود که اشعار پس از انجام پیش‌پردازش‌های لازم، شاخص-گذاری شده و در آن ذخیره می‌شوند. در هنگام جستجو مخزن پرس‌وجوی کاربر را دریافت کرده و اسناد مرتبط با آن را استخراج می‌کند. برای بازیابی اشعار با مضمون خاص به گونه‌ای که کلمات موجود در عبارت جستجو در فاصله کمی نسبت به هم قرار گیرند؛ از پنجره جستجو استفاده شده است. هر پنجره جستجو به اندازه یک مصرع است. بنابراین اشعاری که کلمات موجود در عبارت جستجو در فاصله کمتری از هم قرار بگیرند؛ دارای رتبه بالاتری در نتیجه جستجو خواهند بود.

۴-۱-۴- رابط کاربری

این بخش درخواست‌های جستجوی کاربران را دریافت می‌کند و آن را برای پردازش‌های بعدی وارد بخش پیش‌پردازش پرس‌وجو می‌کند. همچنین در پایان عملیات جستجو نتایج بدست آمده

را به شکل مطلوبی در اختیار کاربر قرار می‌دهد. در جستجوگر صهبا رابط کاربری به صورت یک وب‌سایت طراحی شده است اما می‌توان برای دسترسی بهتر کاربران، نرم‌افزار تلفن‌همراه‌ای نیز برای ارتباط با جستجوگر طراحی کرد.

۴-۱-۵- پیش‌پردازش پرس‌وجو

پرس‌وجوهای کاربران نیز همانند اشعار پیش از هر عملیاتی باید مورد پیش‌پردازش قرار بگیرند. بنابراین در اولین گام عملیات یکسان‌سازی و ریشه‌یابی روی پرس‌وجوی کاربر انجام می‌شود.

۴-۱-۶- خطایاب املایی پرس‌وجو

در اشعار فارسی استفاده از لغات فارسی کهن امری متداول است؛ از طرفی عمده کاربران ممکن است با املای صحیح این لغات آشنا نباشند. بنابراین در جستجوگر شعر نیاز به وجود یک خطایاب املایی برای تصحیح پرس‌وجوهای کاربران دو چندان است.

در جویشگر صهبا تمرکز بر غلط‌های املایی مربوط به کلمات هم‌آوا و هم‌ریخت بوده است. پایه راه حل پیشنهادی برای رفع اینگونه غلط‌ها، الگوریتم فاصله دمرا–لون‌اشتاین[۲۵] است. فاصله لون‌اشتاین میان دو رشته از کمینه تغییرات لازم برای تبدیل یک رشته به دیگری محاسبه می‌شود. این تغییرات شامل درج یک حرف اضافه، حذف یک حرف و یا جایگزینی دو حرف است[11]. این روش می‌تواند بر واژه‌های با طول متفاوت نیز اعمال شود که این تفاوت در طول می‌تواند توسط حذف و یا درج حروف ایجاده شده باشد. برای مثال اگر کاربر کلمه "تحمتن" را جستجو کند؛ با استفاده از الگوریتم گفته شده می‌توان کلمه "تهمتن" را به عنوان شکل صحیح کلمه پیشنهاد کرد (جایگزینی "ه"). فاصله دمرا–لون‌اشتاین این کلمات یک می‌باشد.

۴-۱-۷- اصلاح و بهبود پرس‌وجو

به منظور افزایش دقت و فراخوانی نتایج جستجو دو عملیات روی پرس‌وجوی ورودی انجام می‌شود:

۱.وزن‌دهی لغات پرس‌وجو: چنانچه در پرس‌وجوی کاربر لغات پر کاربردی مانند "از"، "به"، "را" و... وجود داشته باشند به علت پرتکرار بودن این لغات، دقت نتایج بازیابی شده کاهش می‌یابد. از طرفی اگر این لغات را به کلی حذف کنیم در مواردی که کاربر می‌خواهد با جستجو کردن بخشی از یک شعر متن کامل آن را بیابد، دچار مشکل می شویم. برای حل این مشکل در جستجوگر صهبا لغات موجود در پرس‌وجو وزن‌دهی می‌شوند و به لغات پرتکرار وزن کمتری نسبت به بقیه لغات نسبت می‌دهیم. این لغات پرتکرار شامل کلیه افعال فارسی و کلمات خاص پر تکرار در مجموعه اشعار فارسی است.

۲.گسترش پرس‌وجو[۲۶]: در اشعار مختلف از لغات متعددی برای بیان یک مضمون واحد استفاده می‌شود. به عنوان مثال کلمات "می"، "باده"، "ساغر"و "شراب" همگی به مفهوم واحدی اشاره می‌کنند اما صورت‌های ظاهری مختلفی دارند. در این شرایط چنانچه کاربر به دنبال شعری با مضمون "باده" باشد به دلیل تفاوت ظاهری لغات، اشعاری که شامل مترادف-های این لغات باشند را مشاهده نخواهد کرد. برای حل این مشکل در جستجوگر صهبا پیش از ارسال پرس‌وجوهای کاربر به مخزن، پرس‌وجو را گسترش می‌دهیم. برای این‌کار لغات مترادف با لغات موجود در پرس‌وجو را با وزنی کمتر از لغات اصلی به پرس‌وجوی کاربر اضافه می‌کنیم. در نتیجه این‌کار چنانچه خود لغت مد نظر کاربر در شعری وجود داشته باشد آن شعر در نتایج بازیابی شده رتبه بالاتری خواهد داشت و بعد از آن سایر اشعاری که شامل کلمات مترادف بوده‌اند نمایش داده می‌شوند.

۵- نتیجه

امروزه با وجود حضور جستجوگرهای عمومی پیشرفته‌ای همچون گوگل برای جستجو در داده‌های تخصصی نیاز به جستجوگرهای موضوعی احساس می‌شود. در این مقاله مزایای استفاده از جستجوگرهای موضوعی در حوزه‌های تخصصی را بیان کردیم. پس از آن نیاز به یک جستجوگر تخصصی برای اشعار فارسی مطرح شد و چالش‌های پیش‌روی آن را برشمردیم. در بخش چهارم ساختار پیشنهادی برای جستجوگر شعر فارسی و راه‌حل‌های آن برای حل چالش‌های ویژه جستجوی شعر را تشریح کردیم.

سپاسگزاری

این مقاله با اســتفاده از منابع و تجربیات شــرکت کارانس ایرانیان[۲۷] به نگارش درآمده اســت. از همکاری کلیه افراد برای در اختیار قرار دادن منابع و تجربیاتشان متشکریم.

مراجع

[1] Redkar S, Dias N, Laxminarayana JA. "*A survey on domain specific search engine*". International Journal on Advanced Computer Theory and Engineering. 2013.

[2] Almpanidis G, Kotropoulos C, Pitas I. "*Combining text and link analysis for focused crawling—An application for vertical search engines*". Information Systems. 2007.

[3] Steele R. "*Techniques for specialized search engines*". Proc. Internet Computing, Las Vegas. 2001.

[4] Rose, Daniel E., and Danny Levinson. "*Understanding user goals in web search*." In Proceedings of the 13th international conference on World Wide Web, pp. 13-19. ACM, 2004.

[۵] عبداللهی، ص.، جوکار، ع.، "چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب"، مطالعات تربیتی و روانشناسی، دوره دهم، شماره ۲...، دانشگاه فردوسی مشهد، ۱۳۸۸.

[6] Taghva, K., Beckley, R., Sadeh, M., "*A stemming algorithm for the Farsi language*," International Conference on Information Technology: Coding and Computing, pp: 158-162, 2005.

[7] Damerau, F. J. "*A technique for computer detection and correction of spelling errors*". Communications of the ACM. Vol. 7, No. 3, pp: 171–6, 1964

[8] Worthy, J., and Viise, N. M. "*Morphological, phonological, and orthographic differences between the spelling of normally achieving children and basic literacy adults*". Reading and Writing .Vol. 8, No. 2. pp: 139–59, 1996.

[9] Holmes, V. M., and Malone, N. "*Adult spelling strategies*". Reading and Writing. Vol. 17, No. 6, pp: 537–66, 2004.

[10] Kashefi, O., Sharifi, M., & Minaie, B. "*A Novel String Distance Metric for Ranking Persian Respelling Suggestions*". Natural Language Engineering, Vol. 19, No. 2, pp: 259-284, 2013.

[11] Kashefi, O., M. Nasri, and K. Kanani. "*Towards Automatic Persian Spell Checking.*" Tehran, Iran: SCICT (2010).

زیرنویس‌ها

[1] Domain specific search engine
[2] Topical search engine
[3] Vertical search engine
[4] WordNet
[5] Semantic Taxonomy
[6] query
[7] Navigational
[8] Informational
[9] Resource
[10] Normalization
[11] Stemming
[12] Stop words
[13] Index
[14] porter
[15] lovins
[16] precision
[17] recall
[18] Homophone
[19] Homomorph
[20] Single-errors
[21] Multiple-errors
[۲۲] www.sahba.ir
[23] Deterministic Finite-state Machine
[24] Part-of-Speech Tagging
[25] Damerau-Levenshtein Distance
[26] Query expansion
[27] www.karans.co