به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

بررسی موتور جستجوی گنج به منظور بهبود بازیابی اطلاعات در زبان فارسی

تاریخ شروع طرح پژوهشی	۱۳۹۷/۰۶/۱۲	**تاریخ پایان طرح پژوهشی**	۱۳۹۸/۰۷/۲

نام و نام خانوادگی مجری طرح پژوهشی	مرضیه زرین بال ماسوله
نام و نام خانوادگی همکاران طرح پژوهشی	–

دربارهٔ طرح پژوهشی

بازیابی اطلاعات به معنای یافتن محتوایی با طبیعت غیرساخت‌یافته (معمولاً متن) از مجموعه‌هایی به‌منظور برآورده ساختن نیاز(های) اطلاعاتی کاربران است. با بهره‌گیری از روش‌های بازیابی اطلاعات موجود در ادبیات پلتفرم‌های متن‌باز متعددی طراحی شده‌اند. کتابخانه آپاچی لوسن و موتور جستجوی آپاچی سولر از این نوع پلتفرم‌ها هستند. سامانه گنج نیز به منظور جستجو در اسناد و مدارک ذخیره شده در پایگاه داده خود، از این کتابخانه و موتور جستجو به عنوان هسته اصلی فرآیند بازیابی اطلاعات استفاده می‌کند. علی‌رغم کاربردهای گسترده، سولر دارای کاستی‌هایی بوده، تاکنون پژوهشی در زمینه شناسایی و واکاوی این کتابخانه و موتور جستجو در پژوهشگاه انجام نشده و تنها عملکرد رابط کاربری آن و مشکلات موجود بررسی شده است. لذا علی‌رغم ضرورت بررسی کارایی رابط کاربری گنج، تازمانی که هسته اصلی آن (خصوصا با تمرکز بر زبان فارسی) بررسی و اصلاح نشود نمی‌توان انتظار داشت عملکرد سامانه مطلوب باشد. لذا ضروری است تا با بررسی محدودیت‌ها و مشکلات موجود در روش‌های مورد استفاده در موتور جستجوی آپاچی سولر (از منظر بازیابی اطلاعات) دیدگاهی کلان نسبت به آن در پژوهشگاه ایجاد شده تا از این طریق بتوان پیشنهادهایی به منظور ارتقاء عملکرد هسته اصلی فرآیند بازیابی اطلاعات سامانه گنج ارائه داد.

روش پژوهش

پژوهش حاضر در قالب دو مرحله اصلی زیر صورت پذیرفت.

مرحله اول: شناسایی و بررسی کتابخانه آپاچی لوسن و موتور جستجوی آپاچی سولر: در این مرحله کتابخانه آپاچی لوسن و موتور جستجوی آپاچی سولر مطالعه شده و روش‌ها و ابزارهای بکاررفته در هریک از این موارد بررسی شدند.
مرحله دوم: ارائه راهکارها و روش‌های بهبود: شناسایی مشکلات فعلی سامانه گنج خصوصا در زبان فارسی از منظر عملکرد کتابخانه آپاچی لوسن و موتور جستجوی آپاچی سولر و ارائه پیشنهادهایی به منظور بهبود عملکرد بازیابی اطلاعات در زبان فارسی از جمله اقدامات اصلی در این مرحله بودند.

یافته‌های طرح پژوهشی

در این طرح دانش فنی درمورد این موتور جستجو و روش‌های مورد استفاده آن کسب شده و راهکارهایی به منظور ارتقاء نتایج بازیابی در زبان فارسی و در سامانه گنج ارائه شد.

نتیجه‌گیری و پیشنهادها

در طی فرآیند پردازش اسناد و پرس‌جو در سولر اسناد ورودی به پایگاه داده توسط لوسن نمایه شده و ذخیره می‌شوند. درخواست‌های جستجوی کاربر نیز تحلیل شده و نتایج براساس تنظیمات مشخص شده در Solrconfig و Schema بازیابی شده و به کاربر نمایش داده می‌شوند. فرآیند تحلیل متن جزء کلیدی سولر و لوسن بوده و به عنوان ابزاری در زیرسیستم‌های مدیریت اسناد و پردازش پرس‌وجو استفاده می‌شود. سامانه گنج و موتور جستجوی آن نیز از چنین ساختاری برخوردار است.

با بررسی‌های انجام شده می‌توان راهکارهایی را به منظور ارتقاء نتایج بازیابی در زبان فارسی و در سامانه گنج پیشنهاد داد. این راهکارها در قالب دو دسته راهکارهای نیازمند پژوهش و راهکارهای فنی و عملیاتی دسته‌بندی می‌شوند:

راهکارهای نیازمند پژوهش شامل موارد زیر است:

- استفاده از روش‌های جایگزین برای ریشه‌یابی در زبان فارسی و پیاده‌سازی آن در سامانه گنج
- بکارگیری و پیاده‌سازی روش‌های برای واژه–واژه سازی متن و عبارت پرس‌وجو در سامانه گنج
- استفاده از تجزیه‌گر در سامانه گنج
- توسعه و ارتقاء روش بازیابی اطلاعات در سامانه گنج
- فعال‌سازی و پیاده‌سازی کنترل املاء در سامانه گنج

فعال‌سازی و پیاده‌سازی قابلیت‌هایی چون More like this، Facet، Highlight، و قابلیت جستجوی جغرافیایی از جمله راهکارهای فنی و عملیاتی پیشنهادی این طرح پژوهشی است.