

به نام خدا



وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

## خلاصه گزارش طرح پژوهشی سازمانی

### عنوان طرح پژوهشی

طراحی و پیاده سازی وب سرویس استخراج خودکار کلیدواژه از مستندات علمی فارسی

|                                       |                                                        |                        |            |
|---------------------------------------|--------------------------------------------------------|------------------------|------------|
| تاریخ شروع طرح پژوهشی                 | ۱۴۰۲/۱۱/۹                                              | تاریخ پایان طرح پژوهشی | ۱۴۰۳/۱۱/۱۶ |
| نام و نام خانوادگی مجری طرح پژوهشی    | آزاده محبی، امیرحسین صدیقی                             |                        |            |
| نام و نام خانوادگی همکاران طرح پژوهشی | مرضیه زرین بال ماسوله، امیر بادامچی، علی اصغر حجت پناه |                        |            |

### درباره طرح پژوهشی

کلیدواژه های اسناد علمی مانند پارساهای (پایان نامه ها و رساله ها) ثبت شده در پایگاه گنج، بیانگر موضوعات اصلی مطرح شده در سند هستند و در کارهایی نظیر بازیابی اطلاعات، خلاصه سازی خودکار اسناد، خوشه بندی اسناد، طراحی و ساخت آنتولوژی، و دسته بندی موضوعی اسناد، نقش کلیدی دارند. پارساهای کشور در پایگاه گنج ایرانداک ثبت می شوند و طی فرایند نمایه سازی توسط متخصصین کلیدواژه هایی به آنها اختصاص می یابد. تاکنون نزدیک به ۸۰۰ هزار پارسا در پایگاه گنج ثبت شده که حدود ۳۱۵ هزار رکورد نمایه شده اند و بیش از ۴۰۰ هزار رکورد در پایگاه گنج ثبت شده اند اما هنوز کلیدواژه ای به آنها تخصیص نیافته و در صف نمایه سازی هستند. از آنجاییکه فرایند تخصیص کلیدواژه به صورت دستی بسیار زمانبر است، باید از روش های استخراج خودکار کلیدواژه برای اسنادی که تاکنون نمایه نشده اند و آنهایی که از این به بعد در پایگاه گنج ثبت می شوند، استفاده شود. در این طرح پژوهشی با توجه به تجربیات گذشته در طراحی و توسعه روش های استخراج خودکار کلیدواژه در طرح های پژوهشی پیشین که در ایرانداک انجام شده، برای پارساهای ثبت شده که هنوز نمایه نشده اند، کلیدواژه تخصیص یافته است. همچنین سرویسی در قالب API ارائه شده که برای پارساهایی که پس از آن در گنج ثبت می شوند نیز کلیدواژه تخصیص یابد. الگوریتم تخصیص کلیدواژه بر پایه اطلاعات اصطلاحنامه های تخصصی و منابع دانشی مانند فهرست کلیدواژه های کمکی و نیز کلیدواژه های اسناد مشابه است. برای ارزیابی الگوریتم، تعداد محدودی از اسناد (حدود ۱۰۰۰ سند) به صورت تصادفی از حوزه های موضوعی مختلف انتخاب شده و کلیدواژه های آنها براساس الگوریتم تخصیص یافته است. سپس در یک سامانه ارزیابی که در این طرح توسعه یافته، نظرات متخصصین

نمایه‌ساز درباره کلیدواژه‌های تخصیص یافته به آن اسناد دریافت و بررسی شده است. در نهایت از معیارهای دقت و فراخوانی و معیار F برای ارزیابی الگوریتم استفاده شده است. همچنین برای سرویس نمایه‌سازی، API توسعه یافته که با دریافت اطلاعات اولیه سند مانند عنوان و چکیده، در انتها کلیدواژه‌های مناسب را به آن تخصیص می‌دهد. بنابراین اهداف اصلی این طرح عبارتند از: تخصیص کلیدواژه برای پارساهای (پایان‌نامه‌ها و رساله‌های) ثبت شده در گنج که تاکنون نمایه نشده‌اند، و طراحی و توسعه وب سرویس تخصیص خودکار کلیدواژه به مدارکی که از این پس، ثبت می‌شوند.

## روش پژوهش

برای محقق کردن اهداف طرح، الگوریتم استخراج خودکار کلیدواژه، طراحی، پیاده‌سازی شده است و نتایج آن روی مجموعه‌ای از داده‌های سامانه گنج توسط متخصصین و کارشناسان نمایه‌ساز، ارزیابی شده است. الگوریتم پیشنهادی براساس منابع اطلاعاتی زیر، برای هر سند کلیدواژه تخصیص می‌دهد:

- کلیدواژه‌های اسناد مشابهی که تا کنون توسط متخصصین نمایه شده‌اند
  - عنوان، چکیده سند و کلیدواژه‌های نویسنده (اگر موجود است)
  - اصطلاحنامه‌های تخصصی
  - فهرست کلیدواژه‌های کمکی که تا کنون به اسناد اختصاص پیدا کرده‌اند
- با پیاده‌سازی الگوریتم، وب سرویسی طراحی شده که هدف دوم را محقق کند تا براساس آن بتوان برای سایر مدارکی که از این به بعد در سامانه ثبت، وارد می‌شوند نیز به صورت خودکار کلیدواژه تعیین کرد.

## یافته‌های طرح پژوهشی

در این طرح، الگوریتمی ارائه شده که برای یک سند علمی مانند پارساهای گنج، کلیدواژه‌های مناسب تخصیص می‌دهد. این الگوریتم علاوه بر محتوی سند (چکیده و عنوان)، با در نظر گرفتن منابع تخصصی مانند اصطلاحنامه‌ها و فهرست کلیدواژه‌های تخصصی، می‌تواند واژگان و عبارات تخصصی به عنوان کلیدواژه به سند تخصیص دهد. نتایج ارزیابی الگوریتم پیشنهادی و سایر رویکردها نشان داده که تا چه میزان وجود منابع بیرونی مانند اصطلاحنامه‌ها و فهرست واژگان تخصصی می‌تواند در دقت و بازخوانی الگوریتم استخراج کلیدواژه موثر باشد. همچنین با استفاده از الگوریتم طراحی شده، اسناد نمایه نشده در گنج که تا کنون بیش از ۴۰۰ هزار سند هستند، کلیدواژه اختصاص یافته است. نمایه‌سازی این اسناد جهش و تحول بزرگی را در فرایند نمایه‌سازی در ایرانداک ایجاد کرده است. با اتکا به دانش نهاده شده در ایرانداک در طول چند دهه گذشته در فرایند نمایه‌سازی، تلاش‌های موثر در خصوص تدوین اصطلاحنامه‌های تخصصی، و نیز نتایج این طرح و پژوهش‌های پیشین در ایرانداک در زمینه نمایه‌سازی خودکار اسناد علمی، فرایند تخصیص کلیدواژه به اسناد ثبت

شده در ایرانداک هم اکنون می تواند به صورت خودکار انجام شود.

با استفاده از سرویس استخراج خودکار کلیدواژه در این طرح که در قالب یک API در دسترس است هم اکنون می توان برای هر سندی که از این پس ثبت می شود، به صورت خودکار کلیدواژه تخصیص داد.

### نتیجه گیری و پیشنهادها

الگوریتم پیاده سازی شده روی مجموعه ای از داده های گنج اجرا شد و نتایج آن ارزیابی شده است. برای ارزیابی از نظرات متخصصین نمایه ساز در ایرانداک استفاده شده و نظرات ایشان برای کلیدواژه های پیشنهادی هر سند از طریق سامانه ارزیابی توسعه یافته در این طرح، دریافت شده است. نتایج ارزیابی براساس سه شاخص دقت، فراخوانی و معیار F ارائه شده که در نهایت می توان گفت منابع بیرونی مانند اصطلاحنامه ها و فهرست کلیدواژه های کمکی تاثیر بسزایی در بهبود عملکرد الگوریتم پیشنهاد کلیدواژه دارند. همچنین عملکرد الگوریتم پیشنهادی یعنی میزان شاخص دقت، بازخوانی و معیار F از مقدار این شاخص ها در اکثر روش های استخراج خودکار کلیدواژه از مستندات علمی بیشتر یا در همان بازه است و براساس مطالعه انجام شده در پژوهش های پیشین در پژوهشگاه در این زمینه، مقدار این شاخص ها برای مستندات علمی ایرانداک در این الگوریتم در مقایسه با الگوریتم ها بررسی شده در تحقیقات مشابه به میزان بیش از ۹۰ درصد افزایش داشته است.

البته از آنجاییکه در الگوریتم پیشنهادی از اصطلاحنامه و فهرست کلیدواژه های کمکی به عنوان یکی از مهمترین منابع بیرونی برای استخراج کلیدواژه استفاده می شود، در صورتیکه سندی حاوی واژگان تخصصی باشد که در هیچ یک از این منابع نیامده، آنگاه کلیدواژه های پیشنهادی الگوریتم کامل نخواهد بود و احتمالا در بین کلیدواژه های پیشنهادی، آن دسته از واژگان تخصصی دیده نمی شود. برای برطرف کردن این مشکل و ارتقا الگوریتم پیشنهادی، رویکردهایی برای بهبود کلیدواژه های اسنادی که واژگان تخصصی خارج از اصطلاحنامه یا فهرست کلیدواژه های کمکی دارند، در تحقیقات آتی نیاز است که بررسی شوند. همچنین ارائه راهکارهایی برای افزایش زمان اجرای الگوریتم برای هر سند و شناسایی اسنادی که نسبت به بقیه اسناد نمایه شده، کلیدواژه های مناسب یا کاملی ندارند نیز از جمله مسائلی است که باید در تحقیقات آتی بررسی شوند.