

به نام خدا



وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی

تاریخ شروع طرح پژوهشی	۱۳۹۹/۶/۴	تاریخ پایان طرح پژوهشی	۱۴۰۰/۲/۱۸
نام و نام خانوادگی مجری طرح پژوهشی	جلال‌الدین نصیری		
نام و نام خانوادگی همکاران طرح پژوهشی	نصراله پاک‌نیت		

درباره طرح پژوهشی

فراداده داده‌ای است که اطلاعاتی درباره داده‌ای دیگر ارائه می‌دهد. عنوان، نام دانشگاه، رشته تحصیلی، نام دانشجو، استاد راهنما، چکیده و کلمات کلیدی را می‌توان از مهم‌ترین فراداده‌های پایان‌نامه‌ها و رساله‌ها (پارساها) نام برد. از آنجایی که فراداده‌ها در بازیابی اطلاعات نقش منحصر به فردی را ایفا می‌نمایند، استخراج صحیح و ذخیره آن‌ها در سامانه‌های بازیابی اطلاعات حائز اهمیت است. با توجه به حجم بالای متون علمی و سرعت پایین کنترل دستی فراداده در سامانه ثبت پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، همواره نمایه‌سازی پارساهای دریافتی عقب‌تر از پارساهای دریافتی بوده و در نتیجه نمایش اطلاعات در سامانه گنج ایرانداک به روز نیست. با توجه به این مساله، افزایش سرعت کنترل فراداده‌ها از موارد اولویت‌دار سامانه ثبت است.

هدف اصلی این طرح پژوهشی طراحی الگوریتمی است که بتواند فراداده‌ها را از متن کامل پارساها استخراج نماید. چنین الگوریتمی می‌تواند با بررسی سیستمی و هوشمند فراداده‌های پارساها، سرعت نمایه‌سازی پارساها در پایگاه ثبت را افزایش داده و منجر به به‌روز شدن پایگاه گنج گردد.

روش پژوهش

در این پژوهش، روشی برای استخراج فراداده‌ها ارائه شده و تمرکز بر اسناد از نوع پایان‌نامه/رساله در زبان فارسی بوده است. روش پیشنهادی به دو مرحله تشخیص خودکار مرز بدنه و طبقه‌بندی فراداده‌ها تقسیم می‌گردد.

در مرحله نخست، هدف شناسایی پاراگراف‌های بخش پیش‌گفتار پایان‌نامه/رساله است. زیرا فراداده‌های موردنظر از قبیل نام دانشگاه، دانشکده، رشته و مقطع تحصیلی، عنوان، نام دانش‌آموخته، استاد (اساتید) راهنما و مشاور، چکیده و کلمات کلیدی در پیشگفتار هر پایان‌نامه/رساله بیان می‌شوند. بنابراین در مرحله تشخیص مرز بدنه سعی بر این است که شروع فصل اول پایان‌نامه/رساله شناسایی شود. در این مرحله پاراگراف‌ها با استفاده از الگوریتم طبقه‌بندی ماشین بردار پشتیبان به دو کلاس بدنه و فراداده طبقه‌بندی می‌شوند و پس از تشخیص اولین پاراگراف فصل اول پایان‌نامه/رساله، پاراگراف‌های قبل از آن برای استخراج فراداده‌ها به مرحله بعدی هدایت می‌شوند. در فاز بعد - همان‌طور که از نام آن پیداست - هدف طبقه‌بندی پاراگراف‌ها به فراداده‌های ذکر شده است. این طبقه‌بندی نیز با استفاده از ماشین بردار پشتیبان انجام شده است. در نهایت نیز با اعمال پس‌پردازش، فراداده‌های نهایی به دست می‌آیند. در هر مرحله ویژگی‌های متناسب با اسناد پایان‌نامه/رساله فارسی و قالب DOCX تعریف شده است.

در ارزیابی فاز تشخیص بدنه، آزمایش‌هایی برای تنظیم پارامترهای طبقه‌بند ماشین بردار پشتیبان با مقادیر مناسب انجام شده است. اولین آزمایش با هدف یافتن بهترین تابع هسته برای دسته‌بند مذکور نشان می‌دهد که هسته تابع پایه شعاعی نسبت به سایر توابع هسته به درصدهای بهتری دست یافته است. همچنین طی این آزمایش مقادیر مناسب برای مقدار پارامتر کنترل و پارامتر تابع هسته با روش جستجوی شبکه‌ای به دست آمده است. به منظور تنظیم دقیق‌تر پارامترهای کنترل و پارامتر هسته تابع پایه شعاعی در این مرحله، آزمایش دیگری با استفاده از الگوریتم تکاملی خفاش صورت گرفته است. همان‌طور که انتظار می‌رود نتایج به دست آمده در این آزمایش نشان می‌دهد که الگوریتم خفاش می‌تواند سریع‌تر از روش جستجوی شبکه‌ای پارامترهای بهینه (بهینه محلی) را یافته و خطای طبقه‌بندی را کاهش دهد.

یافته‌های طرح پژوهشی

نتایج به دست آمده از تأثیر انواع ویژگی‌ها بر دسته‌بندی پاراگراف‌ها در مرحله تشخیص مرز گویای این است که ویژگی‌های مبتنی بر ذات و طبیعت متن - که در این پژوهش با نام ویژگی‌های هیوریستیک از آن‌ها یاد شد - به تنهایی کفایت نمی‌کند و افزودن ویژگی‌های مبتنی بر توالی پاراگراف‌ها و ویژگی‌های لغوی در بهبود عملکرد طبقه‌بند مؤثر است. در مقابل ویژگی هندسی تعریف شده نه تنها در بهبود درصدهای به دست آمده تأثیری ندارد بلکه ترکیب آن با سایر ویژگی‌ها سبب کاهش یافته‌ها می‌شود. بنابراین به نظر می‌رسد که در مرحله تشخیص مرز می‌توان ویژگی هندسی را کنار گذاشت.

در ارزیابی طبقه‌بندی پاراگراف‌ها در فاز اول، به منظور پیشگیری از بیش‌پردازش از روش اعتبارسنجی متقابل -10

fold استفاده شده است. یافته‌های این آزمون نشان می‌دهد که الگوریتم به طور میانگین با میزان *F1-score* برابر با ۹۵/۶ درصد قادر به دسته‌بندی پاراگراف‌ها است.

در ارزیابی مرحله طبقه‌بندی فراداده‌ها نیز همانند مرحله قبل، آزمایشی برای تعیین پارامترهای الگوریتم ماشین بردار پشتیبان و تابع هسته مناسب انجام شده است. بدین ترتیب در این فاز نیز تابع پایه شعاعی به عنوان تابع هسته الگوریتم ماشین بردار پشتیبان انتخاب شد و پارامترهای تابع هسته و کنترل خطا نیز با استفاده از روش جستجوی شبکه‌ای تنظیم شدند.

بررسی ترکیب ویژگی‌ها بر نحوه عملکرد دسته‌بند فراداده‌ها نشان می‌دهد که ویژگی‌های لغوی به تنهایی برای استخراج فراداده‌ها مناسب نیستند؛ بلکه ترکیب آن‌ها با ویژگی‌های هیوریستیک، قالب، توالی و هندسی در بهبود نتایج تأثیر مثبتی دارند. در این پژوهش ویژگی‌هایی تحت عنوان ویژگی‌های محلی تعریف شدند تا محدودیت پردازش صفحات در فایل‌های DOCX تا حدودی برطرف شود و فراداده‌هایی که بنا به هر دلیلی ممکن است در چند پاراگراف قرار گیرند، بهتر شناسایی شوند. یافته‌های این آزمایش نشان داده است که ادغام ویژگی‌های پیشنهاد شده به ویژگی‌های قبلی نیز موجب افزایش درصدهای به دست آمده شده‌اند.

با انجام اعتبارسنجی متقابل به روش *10-fold* در ارزیابی طبقه‌بند فراداده‌ها، مشخص شد که *F1-score* به طور میانگین برابر با ۹۱/۶ درصد است که نشان از عملکرد نسبتاً مناسب طبقه‌بند دارد.

نتیجه‌گیری و پیشنهادها

محققان در پژوهش‌های آتی می‌توانند در استخراج فراداده‌ها از پایان‌نامه‌ها/رساله‌های فارسی در قالب PDF تمرکز کنند که این موضوع چالش‌های بیشتری نسبت به پژوهش انجام شده دارد. یافتن ارتباط بین فراداده‌ها (مانند اساتید و وابستگی‌های آن‌ها) می‌تواند یکی دیگر از موضوعات پیشنهادی باشد. همچنین با شناسایی خودکار مرز شروع بدنه اصلی پایان‌نامه، می‌توان با کنار گذاشتن پاراگراف‌های پیشگفتار، تنها بدنه اصلی پایان‌نامه/رساله از قبیل متن اصلی، تصاویر، جداول، منابع و سایر بخش‌ها را تحلیل کرد؛ بدون آن که بخش‌های پیشگفتار مورد پردازش قرار گیرد.