به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

ایجاد پایگاه داده از تصاویر موجود در پایگاه اطلاعات علمی ایران (گنج) بر اساس یک روش هوشمند

تاریخ شروع طرح پژوهشی	۹۷/۴/۲۶	**تاریخ پایان طرح پژوهشی**	۹۸/۱۰/۱
نام و نام خانوادگی مجری طرح پژوهشی	آزاده فخرزاده		
نام و نام خانوادگی همکاران طرح پژوهشی	مجتبی زالی		

دربارهٔ طرح پژوهشی

در اسناد و مقالات علمی، تصاویر، حاوی اطلاعات مهمی هستند و در بسیاری از موارد با بررسی آنها به تنهایی می‌توان به ایده اصلی و یا نتایج مهم مقاله علمی پی‌برد، بدون اینکه لازم باشد کل سند را مطالعه کرد. به همین دلیل بسیاری از موتورهای جستجوگر مستندات علمی به دنبال فراهم کردن امکان بازیابی اطلاعات از تصاویر در پایگاه اطلاعاتی خود هستند، به طوری که کاربر با وارد کردن یک جستجو، علاوه بر متن مقالات بتواند به تصاویری هم که به آن جستجو مربوط می‌شود، دسترسی پیدا کند. هم اکنون درپایگاه اطلاعاتی گنج، که حاوی حجم زیادی از مستندات علمی و پایان‌نامه‌ها و رساله‌های فارسی کشور است، امکان جستجو براساس یک عبارت متنی پرس‌وجو و بازیابی و نمایش نتایج جستجو در قالب فراداده‌های متنی (عنوان، چکیده، پدیدآور، سال انتشار،) وجود دارد. لیکن در حال حاضر اطلاعات از تصاویر موجود در اسناد گنج بازیابی نمی‌شود. قدم اول برای بازیابی اطلاعات از تصاویر ایجاد پایگاه داده تصاویر از اسناد است. در این طرح سیستمی خودکار برای ایجاد پایگاه داده از تصاویر موجود در مدارک علمی فارسی در مقیاس بزرگ ارائه می‌شود. سیستم پیشنهادی بخش‌های مختلفی دارد. در مرحله اول باید تصاویر و توضیح متنی آن‌ها استخراج گردد. بر این اساس برای استخراج تصاویر و توضیح متنی آن‌ها یک روش ساختار محور معرفی می‌شود که مبتنی بر چیدمان و آرایش فایل ورد سند است.بدین ترتیب مجموعه‌ای از تصاویر به همراه توضیحات و اطلاعات مربوط به آن‌ها به دست می‌آید که باید در یک پایگاه داده تصاویر با ساختاری

مشخص ذخیره گردند. سپس این اطلاعات برای بازیابی و استفاده‌های آتی در یک موتور جستجو نمایه خواهند شد. در ادامه، روش پیشنهادی در یک مطالعه موردی در پایگاه اطلاعات علمی ایران (گنج) به کار گرفته شد. روش پیشنهادی که با پردازش ساختار و آرایش فایل ورد تصاویر و زیرنویس آن‌ها را استخراج می‌کند در زبان برنامه‌نویسی پایتون پیاده‌سازی شد.

روش پژوهش

روش پژوهش برای بررسی انواع روش‌ها و الگوریتم‌های موجود، روش مطالعات کتابخانه‌ای است. با مطالعه تحقیقات پیشین و بررسی چالش‌های مساله پیش رو، روش جدیدی پیشنهاد شد. سپس روش پیشنهادی بر روی مجموعه‌ای از داده‌ها آزمایش شد تا صحت و دقت، وخطای احتمالی آن بررسی گردد و در صورت نیاز بهبودهایی روی آن صورت گیرد. بنابراین، آماده سازی این مجموعه داده برای انجام آزمایش و اندازه گیری دقت روش هم قسمتی از فعالیت‌های این پژوهش محسوب می‌شود.

گامهای اصلی این پژوهش به شرح زیر است:

۱. مطالعه روش‌های موجود در ادبیات برای استخراج هوشمند تصاویر و توضیح متی آنها ازاسناد علمی
   - بررسی روشهای موجود برای استخراج تصویر و متن مربوطه و چالشهای موجود
   - آماده سازی مجموعه داده‌های آزمایشی برای تست روش پیشنهادی
۲. طراحی و پیاده سازی یک روش استخراج تصاویر، ایجاد پایگاه داده و فراهم ساختن امکان بازیابی تصاویر برای کاربران در هر جستجو
   - طراحی روش مناسب برای استخراج هوشمند تصاویر از اسناد علمی
   - طراحی روش مناسب برای استخراج توضیح متنی از اسناد علمی
   - تست و ارزیابی روش پیشنهادی روی مجموعه داده‌های آزمایشی
   - طراحی و ایجاد پایگاه داده تصاویر، متون و متن‌های توضحی تصاویر
   - طراحی و ایجاد ارتباط بین پایگاه داده محلی و پایگاه داده گنج

به صورت تصادفی ۱۵۰ سند فنی مهندسی با فرمت پی.دی.اف و ورد را از پایگاه داده گنج انتخاب کردیم که اکثر این اسناد پایان‌نامه‌های ارشد و دکتری هستند. و گامهای فوق بر روی این داده ها اعمال شد. نرم افزار تولید شده به زبان پایتون است.

یافته‌های طرح پژوهشی

در طرح حاضر روش‌های استخراج تصویر و زیرنویس جهت ایجاد پایگاه داده تصاویر از اسناد علمی بررسی

شد. برای بررسی کارایی روش پیشنهادی در استخراج تصاویر و توضیحات آن‌ها از یک مطالعه موردی در پایگاه اطلاعات علمی ایران (گنج) کمک گرفتیم. این پایگاه مرجع اصلی دسترسی به تمام متن پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی در ایران می‌باشد. به صورت تصادفی ۱۵۰ سند فنی مهندسی با فرمت پی.دی.اف و ورد را از پایگاه داده گنج انتخاب کردیم که اکثر این اسناد پایان‌نامه‌های ارشد و دکتری هستند. در ایتدا فایل باید تجزیه شود و تصاویر همراه با زیرنویس آنها استخراج شود. از آنجاییکه در ادبیات از فایل پی دی اف استفاده شده است روشهای استخراج تصاویر از فایل پی دی اف هم بررسی شد. به دلیل چالشهای فایل پی دی اف ، از جمله پیروی نکردن آنها ازیک الگوی خاص و تولید تصاویر نامطلوب زیاد در نهایت الگوریتم پیشنهاد شده برای استخراج تصاویر در مورد فایلهای ورد به کار گرفته‌شد. برای استخراج تصاویر و متن متناظر با آن‌ها نیازمند بهره‌برداری از ساختار فایل ورد هستیم، از این رو روش پیشنهادی یک روش ساختار محور مبتنی بر چیدمان و آرایش فایل ورد است. این روش قادر است تصاویر ۴۰ درصد از فایلهارا با دقت ۸۹ درصد استخراج کند. بعد از آنکه تصاویر و اطلاعات مربوطه از فایل استخراج شد، مطابق با مدل داده‌ای که توسط موتور جستجوگر قابل استفاده است ذخیره سازی شد. برای جست‌وجو در تصاویر استخراج شده از اسناد، بخش جست‌وجوی پیشرفته پایگاه با استفاده از زبان برنامه نویسی ruby و چارچوب توسعه نرم‌افزار Ruby on Rails توسعه داده شد. در فصل های بعد خروجی سیستم و نتایج بررسی می شود.

نتیجه‌گیری و پیشنهادها

الگوریتم پیشنهادی در مورد فایلهای ورد با ساختار استاندارد با خطای پایین عمل می کنو و نتیجه مطلوبی دارد. این الگوریتم می تواند می‌تواند بهبود پیدا کند. برخی ازفایلها اگر چه فرمت استاندارد ندارند، اما زیرنویس آنها به شکل شی مجزا در نرم افزار مایکروسافت ورد قابل تشخیص هستند. در مورد این فایلها می‌توان یک نمونه آماری معتبر جمع آوری کرد و چنانچه بین همه آنها از نظر ساختاری شباهت وجود داشته باشد یک الگوریتم قانون محور برای استخراج زیرنویس آنها طراحی کرد. چنانچه در سامانه گنج همراه با فایل سند، لیست زیرنویس تصاویر به صورت اجباری بارگزاری شود، شاید بتوان از این لیست می‌توان برای استخراج بهتر تصاویر بهره برد. در این صورت با روشهای پردازش زبانهای طبیعی و قانون محور موقعیت زیرنویس را در فایل پیدا کرده و ناحیه گرافیکی نزدیک به آن را به آن اختصاص دهیم.

جهت بهبود عملکرد جستجو برای هر تصویر یک جدول برچسب می‌تواند ایجاد شود. برای ایجاد این جدول برچسب می‌توان تمام جملاتی که در آن به تصویر اشاره شده است را استخراج کرد و با استفاده از روشهای پردازش زبانهای طبیعی کلمات کلیدی را استخراج کرد. همچنین می‌توان با استفاده از روشهای پردازش تصویر و نویسه خوانی متن داخل نصویر هم استخراج کرد و به عنوان برچسب در نظر گرفت. کلیه تصاویر استخراج شده می

تواند به یکی از گروههای نمودار خطی، فلوچارت، تصاویر طبیعی، جدول و غیره تقسیم بندی شود و گروه تصویر هم به عنوان یک برچسب در نظر گرفته شود . برای اینکار می‌توان از روشهای یادگیری عمیق استفاده کرد.