

به نام خدا



وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

طراحی یک روش برای برچسب زدن تصاویر استخراج شده از پارساهای موجود در پایگاه اطلاعات علمی ایران (گنج)

تاریخ شروع طرح پژوهشی	۱۳۹۹/۰۲/۱۵	تاریخ پایان طرح پژوهشی	۱۴۰۰/۰۳/۱۶
نام و نام خانوادگی مجری طرح پژوهشی	آزاده فخرزاده		
نام و نام خانوادگی همکاران طرح پژوهشی	امیرحسین صدیقی		

درباره طرح پژوهشی

(بیشینه ۳۰۰ واژه درباره مقدمه، مسئله، و هدف‌های پژوهش و آنچه که انجام، بررسی، یا تعیین شده است.)

در این طرح پژوهشی یک روش جدید برای برچسب‌گذاری تصاویر موجود در متون علمی فارسی معرفی شد. در اسناد 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 پی برد بدون اینکه لازم باشد کل مقاله را مطالعه کرد. به دلیل رشد روز افزون داده‌های تصویری، بسیاری از موتورهای جستجوگر مستندات علمی، به دنبال فراهم کردن امکان بازیابی تصاویر از پایگاه اطلاعاتی خود هستند، به طوری که کاربر با وارد کردن یک جستجو، علاوه بر متن مقالات بتواند به تصاویر مربوط به آن جستجو نیز دسترسی پیدا کند. هم اکنون در پایگاه اطلاعاتی گنج، که حاوی حجم زیادی از مستندات علمی و پایان‌نامه‌ها و رساله‌های فارسی کشور است، امکان جستجو بر اساس یک عبارت متنی و بازیابی و نمایش نتایج جستجو در قالب فراداده‌های متنی (عنوان، چکیده، پدیدآور، سال انتشار) وجود دارد. لیکن تصاویر موجود در اسناد «گنج» بازیابی نمی‌شود. فراهم کردن چنین امکانی در «گنج» به عنوان یک ارزش افزوده می‌تواند آن را از موتورهای جستجوی مشابه متمایز سازد. برای بازیابی تصاویر نیاز است که برچسب‌های توصیف‌کننده‌ای برای

هر تصویر استخراج شود. موتور جستجو با استفاده از این برچسب‌ها می‌تواند تصاویری را بازیابی کند که کلمات جستجو شده در لیست برچسب‌های آنها وجود دارد. در این طرح برای استخراج برچسب تصویر از متن سندی که تصویر به آن تعلق دارد استفاده شده است. زیرنویس و قسمتی از متن سند که در آن، به تصویر مورد نظر اشاره شده است در نظر گرفته شده است. عبارات اسمی در متن با استفاده از پنج روش متفاوت؛ فراوانی عبارات در سند، معکوس فراوانی سند، فراوانی کلمه- معکوس فراوانی سند، شباهت کسینوسی عبارات با زیرنویس و ترکیب روش فراوانی کلمه- معکوس فراوانی سند و شباهت کسینوسی با زیرنویس، رتبه‌بندی می‌شوند. برچسب‌های انتخابی برای تصویر در هر روش، عبارات اسمی با رتبه بالاتر در آن روش است. روش‌های معرفی شده با استفاده از داده آزمایشی از گنج، ارزیابی شدند. طبق نتایج بدست آمده بر روی داده آزمایشی در این تحقیق، روش فراوانی کلمه- معکوس فراوانی سند با معیار F_1 ، ۳۸٪ بهترین روش برای برچسب زدن تصاویر و پس از آن ترکیب روش فراوانی کلمه- معکوس فراوانی سند و شباهت کسینوسی با زیرنویس با معیار F_1 ، ۳۶٪ قرار دارد. از آنجاییکه در مواردی زیرنویس تصاویر توصیف کاملی از تصاویر نمی‌دهد، با در نظر گرفتن شباهت کسینوسی با زیرنویس تصاویر، معیار F_1 کاهش می‌یابد.

روش پژوهش

(بیشینه ۳۰۰ واژه درباره گام‌های پژوهش و چگونگی گردآوری و تحلیل داده‌ها (ابزار، جامعه، نمونه، روش نمونه‌گیری، روش‌های تجزیه و تحلیل داده‌ها).

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. روش پیشنهادی بر روی مجموعه‌ای از داده‌ها آزمایش شد تا صحت و دقت، و خطای احتمالی آن بررسی گردد و در صورت نیاز بهبودهایی روی آن صورت گیرد.

در گام‌های اصلی این پژوهش موارد زیر بررسی و انجام شد

۱) مطالعه روش‌های موجود در زمینه استخراج برچسب تصاویر و آماده‌سازی داده آزمایشی

- بررسی روش‌های موجود در ادبیات برای استخراج برچسب تصاویر و چالش‌های موجود
- بررسی روش‌های استخراج کلمات کلیدی متن

- آماده‌سازی داده آزمایشی برای انجام آزمایشات:

برای انجام این تحقیق، ۴۰۰ فایل ورد، پایان نامه و رساله به صورت تصادفی از پایگاه داده گنج انتخاب شد. از این تعداد، ۳۲۲ فایل محتوای صحیح داشتند و برای انجام آزمایش در نظر گرفته شد. کد استخراج تصویر از فایل ورد، قادر به استخراج تصاویر از فایل‌هایی است که فرمت استاندارد ورد را دارند. به دلیل اینکه تعدادی از فایل‌ها خالی از تصویر بودند و همچنین تعداد از آنها فرمت استاندارد نداشتند از مجموع ۳۲۲ فایل، از ۹۵ فایل تصاویر استخراج شد. از مجموعه تصاویر استخراج شده ۱۵۰ تصویر به صورت تصادفی انتخاب شده و بر اساس محتوای زیرنویس و متن مرجع توسط سه خبره مستقل در حوزه مهندسی وفنی برچسب موضوعی خورده‌اند. این ۱۵۰ تصویر برای محاسبه خطای سیستم استفاده می‌شود

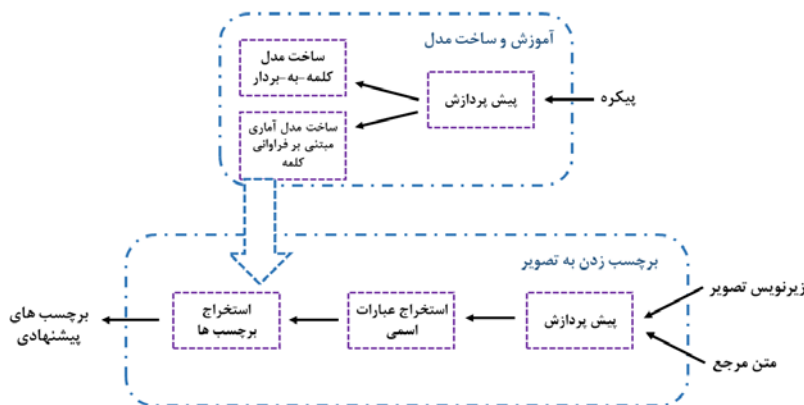
(۲) طراحی یک روش برای برچسب زدن تصاویر استخراج شده از اسناد علمی موجود در گنج

- استخراج برچسب‌های تصویر با استفاده از زیرنویس آن و متن سند مربوطه
- آزمایش و ارزیابی روش پیشنهادی روی مجموعه داده‌های آزمایشی

یافته‌های طرح پژوهشی

(بیشینه ۵۰۰ واژه درباره یافته‌ها و دستاوردهای کلیدی پژوهش در پاسخ به پرسش‌ها یا بررسی فرضیه‌های پژوهش).

شکل زیر، واحدهای مختلف سیستم طراحی شده را نشان می‌دهد. این سیستم با تجزیه و تحلیل متن همراه تصویر با استفاده از روش‌های پردازش زبان طبیعی، برچسب‌های توصیف کننده تصویر را استخراج می‌کند.



از متن همراه تصویر که شامل زیرنویس و متن مرجع است، عبارات اسمی که شامل همه ترکیبات unigram، bigrams و trigram متن مرجع است، استخراج شد. سپس عبارات اسمی استخراج شده با استفاده از پنج روش فراوانی کلمه، معکوس فراوانی سند، فراوانی کلمه- معکوس فراوانی سند، شباهت کسینوسی عبارات با زیرنویس و ترکیب روش فراوانی کلمه- معکوس فراوانی سند و شباهت کسینوسی رتبه بندی شدند.

هرچه تعداد برجسب‌های انتخابی توسط سیستم را افزایش دهیم بازخوانی سیستم بهتر می‌شود اما به دلیل ظاهر شدن مثبت‌های اشتباه، دقت سیستم و به همان نسبت معیار F_1 کاهش پیدا می‌کند. در هر روش تعدادی از برجسب‌های برتر در نظر گرفته می‌شود که میانگین معیار F_1 به ازای آن بیشینه است. بر اساس مشاهدات ما، روش‌های فراوانی کلمه و معکوس فراوانی سند عملکرد ضعیفی دارند و روش‌های شباهت کسینوسی، فراوانی کلمه- معکوس فراوانی سند و ترکیب شباهت کسینوسی با ، فراوانی کلمه- معکوس فراوانی سند خروجی قابل مقایسه‌ای دارند. از بین همه روشها ، فراوانی کلمه- معکوس فراوانی سند بیشترین معیار F_1 را دارد و به عنوان روش مطلوب معرفی می‌شود.

عباراتی که ، فراوانی کلمه- معکوس فراوانی سند بالاتری دارند در واقع کلمات کلیدی متن هستند. انتظار اولیه ما این بود که کلمات موجود در زیرنویس بهترین توصیف کننده‌های تصویر باشند و بر این اساس با رتبه بندی مجدد کلمات کلیدی متن بر اساس شباهت کسینوسی با زیرنویس برجسب‌های بهتری برای تصویر به دست بیاید. اما نتایج آزمایشات نشان داد در نظر گرفتن شباهت کسینوسی با زیرنویس، اگر چه نتایج قابل مقایسه‌ای با فراوانی کلمه- معکوس فراوانی سند دارد، باعث بهبود خروجی آن نشد. دلیل اصلی این امر این است که در خیلی از موارد زیرنویس‌های انتخاب شده توسط نویسنده سند دقیق نیستند.

نتیجه‌گیری و پیشنهادها

(بیشینه ۳۰۰ واژه بحث و نتیجه‌گیری و بیان پیشنهاد‌های برگرفته از پژوهش).

روشی برای برجسب زدن تصاویر موجود در اسناد علمی معرفی شد. تصاویر حاوی اطلاعات مهمی از اسناد علمی هستند، به همین دلیل امروزه تحقیقات در زمینه بازیابی تصاویر از اسناد علمی، بسیار رو به رشد است. به دلیل پیچیدگی‌های محتوایی تصاویر علمی، در این تحقیق از زیرنویس و قسمتی از متن سند که شامل اشاره به تصویر است، برای برجسب زدن تصاویر استفاده شد. آزمایشات نشان داد انتخاب برجسبها با استفاده از روش فراوانی کلمه- معکوس فراوانی سند عملکرد بهتری نسبت سایر روش‌های پیشنهادی دارد.

یکی از مشکلات روش پیشنهادی و البته تمام روش‌های مشابه که بر اساس مدل‌سازی آماری طراحی شده است این

است که در خروجی نهایی عبارات مثبت اشتباهی مشاهده می‌شود که در خیلی موارد بارمعنایی کمی دارند. در صورت دسترسی به پیکره عظیم داده در حوزه تخصصی مورد نظر، می‌توانیم مدل‌های کلماتی قابل اعتمادتری بسازیم. مدل کلماتی بهتر خطای سیستم را کاهش می‌دهد. همچنین دسترسی به یک بانک داده عظیم از واژه‌های کلیدی موجود در حوزه مربوطه می‌تواند در فیلتر کردن عبارات بی‌معنی کمک کننده باشد.

در روش ارائه شده، با بررسی متنی که به تصویر اشاره می‌کند برچسب‌های تصویر استخراج شده‌است و محتوای خود تصویر در نظر گرفته نشده است. از جمله پیشنهادها برای تحقیقات آتی این است که محتوای خود تصویر هم بررسی شود. از آنجایی که تنوع تصاویر در اسناد علمی زیاد است و ویژگی‌های یک گروه با گروه دیگر از تصاویر خیلی متفاوت است، بنابراین در قدم اول باید تصاویر را گروه‌بندی کرد و سپس با استفاده از روش‌های پردازش تصویر اطلاعات مناسب در خود تصاویر را برای برچسب زدن در نظر گرفت.