به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی

عنوان طرح پژوهشی

ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی براساس سیستم‌های پیشنهاددهنده

تاریخ شروع طرح پژوهشی	۱۳۹۶/۳/۱۰	**تاریخ پایان طرح پژوهشی**	۱۳۹۷/۱۲/۱۸

نام و نام خانوادگی مجری طرح پژوهشی	آزاده محبی
نام و نام خانوادگی همکاران طرح پژوهشی	همکار اصلی: عمار جلالی‌منش گروه تحقیقاتی: امیر بادامچی، فرنوش بیات‌ماکو

دربارهٔ طرح پژوهشی

کلیدواژه‌های یک سند، توصیفگرهای مفهومی هستند که می‌توانند در جستجو و بازیابی اطلاعات و نیز اشاعه آنها بکارگرفته شوند. در پایگاه‌های دربردارنده اسناد علمی مانند پایگاه علمی گنج ایرانداک، کلیدواژه‌ها نقش مهمتری دارند و تخصیص کلیدواژه‌های تخصصی نیز چالش‌برانگیزتر خواهد بود، زیرا در این پایگاه‌ها اسناد تخصصی با حوزه‌های علمی مختلفی وجود دارند. با توجه به افزایش حجم تولید و ثبت مستندات علمی، نیاز است که فرایند تخصیص کلیدواژه با سرعت بیشتری صورت گیرد و از روش‌های ماشینی هوشمند نیز استفاده گردد. در بسیاری از پایگاه‌های اطلاعات علمی دنیا از روش‌های خودکار برای استخراج کلیدواژه‌ها استفاده می‌شود. تعدادی از این روش‌ها بر مبنای تحلیل آماری متون و استفاده از روش‌های یادگیری ماشین هستند، تعدادی بر مبنای تحلیل معنایی متون به واسطه اصطلاح‌نامه‌های تخصصی و هستان‌شناسی، و در تعدادی دیگر از این روش‌ها از تلفیق هر دو استفاده می‌شود. بر همین اساس، در این طرح پژوهشی روشی برای پیشنهاد کلیدواژه به مستندات علمی فارسی ارائه شده که بر مبنای روش‌های هوشمند پردازش متن و یادگیری ماشین عمل می‌کند. روش پیشنهادی بر مبنای سیستم‌های پیشنهاددهنده و استدلال نمونه‌محور طراحی شده که براساس آن، مجموعه‌ای از کلیدواژه‌های مرتبط با یک سند پیشنهاد می‌شود. روش پیشنهادی براساس استدلال نمونه محور عمل می‌کند که در آن فرض بر این است که اسناد مشابه کلیدواژه‌های مشابه دارند. بر همین اساس، ابتدا اسناد مشابه با یک سند جدید براساس روش‌های TFIDF و روش‌های بازنمایی کلمه-به-بردار(Word2Vec)، بازیابی می‌شوند. سپس کلیدواژه‌های کاندید از بین کلیدواژه‌های اسناد مشابه بازیابی شده درنظر گرفته می‌شوند. در نهایت براساس تابع رتبه‌بندی پیشنهادی، کلیدواژه‌های مناسب از بین فهرست کلیدواژه‌های کاندید انتخاب می‌شوند. روش پیشنهادی بر روی مجموعه‌ای از اسناد پایگاه گنج در سه حوزه فنی و مهندسی، هنر و ادبیات، و علوم انسانی، پیاده‌سازی شده و نتایج آن با معیارهایی نظیر دقت، فراخوانی و نظرات متخصصین ارزیابی شده است.

روش پژوهش

روش پیشنهادی بر مبنای عملکرد سیستم‌های پیشنهاددهنده محتوی محور عمل می‌کند وشیوه عملکرد آن برای پیشنهاد کلیدواژه براساس منطق استدلال نمونه‌محور است. در استدلال نمونه‌محور فرض بر آن است که مسائل شمابه، راه‌حل‌های مشابهی دارند. بر همین اساس، در روش پیشنهادی برای یک سند جدید، اسناد مشابه با آن ، بازیابی می‌شوند و کلیدواژه‌های سند جدید از بین کلیدواژه‌های اسناد مشابه براساس یک تابع امتیازدهی، انتخاب می‌شوند. دو بخش اصلی در روش پیشنهادی وجود دارد: بخش آموزش برای ساخت مدل و بخش بازیابی و پیشنهاد کلیدواژه.

روش پیشنهادی برای تعدادی از پایان‌نامه‌های ارشد و دکتری تعدادی از دانشگاه‌های ایران پیاده‌سازی شده است. این پایان‌نامه‌ها از پایگاه اطلاعات گنج متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) در سه حوزه موضوعی هنر و ادبیت، فنی و مهندسی، و علوم انسانی استخراج شده است.

یافته‌های طرح پژوهشی

برای ارزیابی روش پیشنهادی از دو رویکرد استفاده شده است: رویکرد ارزیابی ماشینی براساس داده‌های آموزشی و آزمایشی، رویکرد نظر متخصصین. در هر دو رویکرد داده‌ها به سه مجموعه آموزش، اعتبارسنجی، و آزمایش تقسیم شده است. داده‌های آموزش برای آموزش مدل یادگیری، داده‌های اعتبارسنجی برای تنظیم پارامترهای مدل و داده‌های آزمایشی برای ارزیابی عملکرد روش پیشنهادی مورد استفاده قرار گرفته است. در روش ارزیابی ماشینی براساس داده‌های آزمایشی، و کلیدواژه‌های تخصصی یافته به اسناد، میزان، دقت و بازخوانی اندازه‌گیری شده است. در روش ارزیابی براساس نظر متخصص، سامانه ارزیابی طراحی شده که متخصصین حوزه نمایه‌سازی، کلیدواژه‌های پیشنهادی را از نظر میزان ارتباط آنها با هر سند ارزیابی نموده‌اند و نتایج نیز براساس شاخص دقت بررسی شده است.

همچنین برای بررسی اثرگذاری پارامترهای مختلف در روش پیشنهادی، آزمایش‌های مختلفی انجام شده است و عملکرد آن براساس شاخص بازخوانی و شباهت بین کلیدواژه‌های پیشنهادی و سند، بررسی شده است. براساس آزمایش‌های انجام شده، می‌توان گفت که آموزش مدل‌های جداگانه برای هر حوزه موضوعی بر روی مجموعه داده‌های موضوعی، می‌تواند نتایج بهتری را ارائه دهد. همچنین از آنجاییکه نتایج روش پیشنهادی وابسته به تعداد اسناد مشابه بازیابی شده ، بنابراین الگوریتم بازیابی نقش مهمی را در عملکرد روش پیشنهادی ایفا می‌نماید.

بررسی شباهت معنایی بین کلیدواژه‌های پیشنهادی با سند و شباهت معنایی بین کلیدواژه‌های اصلی با سند، نشان می‌دهد که اختلاف معناداری بین شباهت‌ها وجود ندارد و بر همین اساس می‌توان گفت که از نظر معنایی کلیدواژه‌های پیشنهادی با سند از نظر معنایی مرتبط هستند.

نتیجه‌گیری و پیشنهادها

از آنجاییکه کلیدواژه‌های پیشنهاد شده در کنار دانش نمایه‌ساز یا سایر روش‌های استخراج کلیدواژه که بر اساس استخراج واژگان مهم از متن کامل سند عمل می‌کنند، می‌تواند بکار گرفته شود، روش پیشنهادی می‌تواند به عنوان یکی از منابع دانشی در فرایند نمایه‌سازی درنظر گرفته شود.

یکی از پیشنهادها برای تحقیقات آتی اینست که روش‌های استخراج کلیدواژه از متن که بر مبنای اطلاعات آماری واژگان کاندید درون متن عمل می‌کنند در کنار روش پیشنهادی در نظر گرفته شود.

علاوه بر آن، از آنجاییکه عملکرد روش‌های مدلسازی مانند کلمه-به-بردار وابسته به حجم داده‌های آموزش است، برای بهبود عملکرد روش پیشنهادی می‌توان از مجموعه داده بزرگ تری با داده‌های موضوعی متنوع تر استفاده نمود.

یکی دیگر از پیشنهادها اینست که با افزایش مجموعه داده‌ها، حوزه‌های موضوعی به صورت تخصصی‌تر و با تعداد بیشتر مشخص گردد و در فرایند بازیابی اسناد مشابه، به پایگاه داده موضوعی تخصصی‌تری مراجعه شود.