به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش برای پژوهشگران

شناسایی مسائل و ارائهٔ راهکارهای زبانشناختی سازمان‌دهی و بازیابی اطلاعات در سامانهٔ اطلاعات علمی ایران (گنج)

1397

مجری: زهرا محمودزاده

همکار: زهرا دهسرایی

معرفی پژوهش

پردازش متن یکی از چالش‌هایی است که سیستم‌های بازیابی اطلاعات با آن روبرو هستند. ساختار زبانی متن شامل ویژگی‌های ساختواژی-نحوی-معنایی و لغوی متن تأثیر بسزایی بر دقت بازیابی دارد. در سامانهٔ گنج به عنوان مثال، مسائل خط فارسی از جمله عواملی است که سبب کاهش دقت بازیابی می‌شود. در این طرح هدف بر آن بوده است تا حد امکان مسائل بازیابی و سامان‌دهی اطلاعات به ویژه مسائلی که ریشه در پردازش متن و ساختار زبانی آن دارد، بر اساس ادبیات موجود شناسایی شده. پس از مطالعه و دسته‌بندی آنها، راه‌کارهای لازم بر اساس ادبیات موجود ارائه شود.

۱- مرحله سازمان‌دهی و نمایه‌سازی

در این مرحله نمایه‌ساز اطلاعات پایان‌نامه ها را وارد می‌کند. شناسایی مسائل نمایه‌سازی صورت‌های زبانی و ارائه راهکار برای حل یا بهبود آنها براساس ادبیات موجود از اهداف این مرحله است. سامانهٔ اطلاعات علمی ایران (گنج) حاوی داده‌های متنی از نوع پایان‌نامه است. اطلاعات پایان‌نامه‌ها طی فرایند نمایه‌سازی (سازمان‌دهی) به صورت‌های زیر و با دو زبان فارسی و انگلیسی ذخیره می‌شود: چیکده (فارسی –انگیسی)، عنوان (فارسی – انگیسی)، کلیدواژه (فارسی – انگیسی)، پدیدآوران شخص (پدیدآور، استاد راهنما، استاد مشاور)، پدیدآوران سازمان (دانشگاه، دانشکده)، اطلاعات تخصصی پایان‌نامه (رشته/گرایش، مکان تولید)، حوزهٔ موضوعی (در دو سطح).

۲- مرحله بازیابی

بنابراین در فرایند بازیابی که بر روی داده‌های سامان‌دهی شده صورت می‌گیرد، می‌توان اطلاعات داده‌ها (یا پایان‌نامه‌ها) را براساس هفت صورت بالا بدست آورد. اما عوامل متعددی برخواسته از ساختار زبانی صورت‌های سامان‌دهی و ذخیره شده شامل ویژگی‌های نحوی، معنایی، صرفی،

لغوی و املایی بر دقت بازیابی تأثیر گذاشته، آن را کاهش می‌دهد. کاربر برای بازیابی اطلاعات ذخیره شده به وسیلهٔ نمایه‌ساز با موانع، خطاها و کم و کاستی‌هایی روبرو است. شناسایی این موانع و ارائه راهکار بر اساس ادبیات موجود هدف دیگر این طرح است.

روش پژوهش

گام اول: شناسایی مسائل زبانشناختی سازماندهی و بازیابی اطلاعات سامانه اطلاعات علمی ایران به روش مطالعه و مصاحبه و جمع اوری میدانی نمونه ها از پایگاه گنج. مشاهده گزارش‌های موجود در این زمینه، مطالعه ادبیات موجود در این زمینه، مصاحبه با نمایه‌سازان مرکز ایرانداک و بخش بازیابی اطلاعات و جمع‌اوری داده ها از پایگاه گنج.

گام دوم: ارائه راهکار برای مسائل بالا به روش کتابخانه‌ای: مطالعه ادبیات موجود

یافته‌ها

مسائل و راهکارهای ارائه شده در جدول های زیر به صورت خلاصه ارائه شده است. بدیهی است مسائل ساماندهی در بازیابی نیز ملاحظه می شود.

مسائل ساماندهی	راهکار
ترجمه به فارسی یا انگلیسی اطلاعات کتابشناختی	۱- بهره گیری از زبانشناسان و متخصصان زبان فارسی ۲- افزایش قابلیت نرم افزار نمایه سازی ۳- تشکیل گروه تصمیم ساز متشکل از زبانشناس، متخصص حوزه های علمی و نمایه ساز
استفاده از معادل‌های متعدد برای یک واژه لاتین	
تعیین معادل فارسی کلیدواژه‌های لاتین	
تعیین اختصار فارسی برای یک واژه لاتین	
نوشتن صورت های مخفف لاتین به اشکال مختلف	
تعدد صورت های نوشتاری اسامی خاص افراد (پدیدآوران)	نرم افزارهای خطایاب زبان فارسی موجود در بازار کمک چندانی به حل مسائل ناشی از خطاهای مختلف نگارشی و املایی و نیز درج فرمولها نمی کنند. هر چند می توانند مقداری از مسائل و خطاها را به نمایه ساز نشان دهند.
انواع خطاهای املایی و نگارشی در هر دو سطح فارسی و لاتین	
مسائل مربوط به درج علائم ریاضی و فرمولهای شیمی در چکیده و عنوان	
مطالب نامربوط در چکیده	

مسائل بازیابی	راهکار
صورت‌های متفاوت نوشتاری ۱- ناشی از کاربرد علامتهای مختلف نوشتاری: تشدید، مساد، همزه، تنوین، آوا، نقطه، خط تیره ۲- صورت‌های متفاوت کسرهٔ اضافه ۳- صورت‌های مختلف جمع ۴- صورت‌های مختلف اختصارات	بررسی ادبیات موجود نشان می دهد که با کاربرد روش ها و تکنیک های پردازش زبان طبیعی در مدلهای بازیابی می تواند دقت بازیابی را افزایش داد. بنابراین با مشخص کردن روش بازیابی اطلاعات در سامانه گنج و سپس افزودن تکنیک های NLP به مدل های موجود می توان دقت بازیابی در این سامانه را بهبود بخشید.
مسائل معنایی-صرفی- نحوی ۱-کلمات چندمعنا با املای یکسان (اما صورت آوایی یکسان یا متفاوت) ۲- صورت های متفاوت واژگانی با معنای یکسان یا مرتبط ۳- صورت‌های کوتاه / کامل اسامی یا سازمان‌ها	

نتیجه گیری و پیشنهادها

در این پژوهش ابتدا با بررسی میدانی پایگاه اطلاعات علمی ایران (گنج)، مسائل زبانشناختی بازیابی و ساماندهی اطلاعات در این پایگاه بدست آمد و سپس راهکارهایی ارائه شد. مسائل نمایه سازی شامل موارد زیر است: ترجمه به فارسی یا انگلیسی اطلاعات کتابشناختی، استفاده از معادل‌های متعدد برای یک واژه لاتین، یافتن معادل فارسی کلیدواژه‌های لاتین، تعیین اختصار فارسی برای یک واژه لاتین، نوشتن صورت های مخفف لاتین به اشکال مختلف، انواع خطاهای املایی و نگارشی در هر دو سطح فارسی و لاتین، مسائل مربوط به درج علائم ریاضی و فرمولهای شیمی در چکیده و عنوان، مطالب نامربوط در چکیده. برای حل مسائل زبانشناختی ساماندهی و نمایه سازی به نظر می رسد علاوه بر بهبود قابلیت نرم افزار نمایه سازی، لازم است دانش تخصصی در زمینه علوم مختلف و نیز زبان فارسی و زبانشناسی بکار گرفته شود.

مسائل بازیابی به دو دسته تقسیم می شوند الف) صورت های متفاوت نوشتاری ناشی از کاربرد علامتهای مختلف نوشتاری: تشدید، همزه، تنوین،آوا، نقطه، خط تیره، صورت‌های متفاوت کسرهٔ اضافه و صورت‌های مختلف جمع و صورت‌های مختلف اختصارات و ب) صورت های متفاوت ناشی از مسائل معنایی، صرفی و نحوی: کلمات چندمعنا با املای یکسان (اما صورت آوایی یکسان یا متفاوت)، صورت های متفاوت واژگانی با معنای یکسان یا مرتبط و صورت‌های کوتاه / کامل اسامی یا سازمان‌ها. بررسی ادبیات مربوط به بازیابی اطلاعات نشان می دهد که بکار گیری روشها و تکنیک های پردازش زبان طبیعی تأثیر معناداری در بهبود دقت و بازخوانی سیستم های بازیابی اطلاعات دارد.