

به نام خدا



وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

ارائه یک ماژول نرم‌افزاری در راستای ایجاد شبکه استنادی پارساهای موجود در پایگاه اطلاعات علمی ایران (گنج)

تاریخ شروع طرح پژوهشی	۱۳۹۸/۵/۶	تاریخ پایان طرح پژوهشی	۱۳۹۹/۷/۲۸
نام و نام خانوادگی مجری طرح پژوهشی	نصراله پاک‌نیت		
نام و نام خانوادگی همکاران طرح پژوهشی	جلال‌الدین نصیری		

درباره طرح پژوهشی

ارجاعات، که در مدارک علمی به وسیله رشته‌های مرجع در انتهای مدارک علمی آورده می‌شوند، نقش مهمی در کتابخانه‌های دیجیتالی انتشار علمی ایفا می‌کنند. علاوه بر استفاده از ارجاعات به عنوان ابزاری برای یافتن اطلاعات موردعلاقه، از آن‌ها به عنوان معیاری برای بررسی تأثیر و اهمیت مدارک و در سطحی بالاتر، برای بررسی صلاحیت پژوهشگران برای ارتقاء، اعطای پژوهانه و پروژه‌های تحقیقاتی استفاده می‌شود. استفاده از ارجاعات برای اهداف موردنظر نیازمند ساخت شبکه تحلیل استنادی است تا روابط بین موجودیت‌هایی مانند اسناد، پژوهشگران، نشریات و ... مشخص گردد. برای ساخت شبکه تحلیل استنادی بر روی مجموعه‌ای از اسناد علمی، باید مراحل زیر انجام شود:

۱. استخراج فراداده‌های متناظر با هر سند ورودی.

۲. استخراج بخش مراجع هر سند و تقسیم آن به رشته‌های مرجع مجزا.

۳. استخراج فراداده‌هایی مانند اسامی پژوهشگران، عنوان، محل انتشار، سال و ... از هر رشته مرجع.

۴. به‌کارگیری الگوریتم‌های تطبیق رشته برای بررسی مطابقت رشته‌های مرجع و مولفه‌های آن‌ها.

انجام مراحل فوق توسط کاربر انسانی به راحتی قابل انجام است، اما ماشینی‌سازی این مراحل دارای چالش‌هایی

اساسی است. بررسی‌های صورت گرفته نشان می‌دهد که از یک طرف ساخت شبکه تحلیل استنادی وابسته به زبان بوده و از طرف دیگر، تاکنون روشی برای ساخت شبکه تحلیل استنادی در زبان فارسی ارائه نشده است.

با توجه به نقش پژوهشگاه علوم و فناوری اطلاعات ایران در نگهداری، نمایه‌سازی و ارائه دانش افزوده از مدارک علمی کشور، در این پژوهش به مسأله ساخت خودکار شبکه تحلیل استنادی بر روی مجموعه پارساهای پایگاه (گنج) پرداخته شده است. در این راستا، با در نظر گرفتن روش‌های موجود برای ساخت خودکار شبکه تحلیل استنادی در زبان انگلیسی و همچنین ویژگی‌های زبان فارسی و مراجع در این زبان، الگوریتم‌های موردنیاز برای ساخت خودکار شبکه تحلیل استنادی ایجاد و تحت یک ماژول نرم‌افزاری ارائه شده است.

روش پژوهش

این پژوهش در دو گام صورت گرفته است. در گام نخست، در ابتدا با رویکردی کتابخانه‌ای به بررسی کارهای انجام شده بر روی مسائل موجود در حیطه این پژوهش در منابع فارسی و انگلیسی پرداخته شده و در ادامه، با استفاده از نتایج به دست آمده، الگوریتم‌های لازم برای ارائه ماژول نرم‌افزاری موردنظر طراحی شده است. الگوریتم‌های ارائه شده در این گام مبتنی بر روش‌های یادگیری ماشینی، به طور مشخص روش ماشین بردار پشتیبان، و الگوریتم‌های تطبیق رشته است. پیاده‌سازی آزمایشی الگوریتم‌های ارائه شده با استفاده از زبان برنامه‌نویسی پایتون انجام شده و با استفاده از معیارهای دقت، فراخوانی و اف-۱ کیفیت الگوریتم‌ها بررسی شده است. لازم به ذکر است که بررسی کیفیت صورت گرفته بر روی مجموعه داده‌هایی ایجاد شده در این پژوهش که با استفاده از مجموعه‌ای تصادفی از پارساهای مربوطه به بازه زمانی ۱۳۹۳ تا ۱۳۹۸ پایگاه گنج ایجاد شده، انجام شده است. در گام دوم، یکپارچه‌سازی و تجمیع الگوریتم‌های ارائه شده با استفاده از ابزارهای موردنظر پژوهشگاه (زبان برنامه‌نویسی ++C، پایگاه داده MySQL و ...) صورت گرفته و نتیجه تحت یک ماژول نرم‌افزاری ارائه شده است.

یافته‌های طرح پژوهشی

در این پژوهش، به مساله ساخت خودکار شبکه تحلیل استنادی بر روی مجموعه پارساهای پایگاه گنج پرداخته شد. در ابتدا پیشینه مساله و کارهای صورت گرفته برای تجزیه رشته‌های مرجع در زبان انگلیسی بررسی شد. در ادامه، از روش یادگیری ماشینی ماشین بردار پشتیبان و الگوریتم‌های تطبیق رشته تقریبی استفاده گشته و الگوریتم‌هایی جهت حل مسائل موجود در مساله اصلی پژوهش یعنی استخراج بخش مراجع هر سند و تقسیم آن به رشته‌های مرجع

مجزا، تعیین نوع رشته‌های مرجع، استخراج فراداده‌های هر رشته مرجع و بررسی مطابقت رشته‌های مرجع و مولفه‌های آن‌ها ارائه شد. راه کارهای ارائه شده برای سه مساله اول مبتنی بر روش ماشین بردار پشتیبان است که یک روش یادگیری ماشینی با نظارت است. نتایج پیاده‌سازی و آزمایش الگوریتم‌های ارائه شده با استفاده از معیارهای موجود برای بررسی کیفیت این گونه روش‌ها (یعنی معیارهای دقت، فراخوانی و اف-۱) نشانگر کیفیت بالای روش‌های پیشنهادی در حل مسائل موردنظر است. در ادامه، یکپارچه‌سازی و پیاده‌سازی نهایی روش پیشنهادی با توجه به ویژگی‌های موردنیاز پژوهشگاه صورت گرفت و آزمایشاتی نهایی برای بررسی کیفیت کلی روش ارائه شده انجام شد. مقادیر نهایی به دست آمده در ارزیابی روش پیشنهادی برابر با مقدار ۰,۹۹ برای معیار دقت، ۰,۷۳ برای معیار فراخوانی و ۰,۸۵ برای معیار اف-۱ می‌باشد که نشان‌دهنده کیفیت مناسب روش پیشنهادی است.

نتیجه‌گیری و پیشنهادها

همانطور که در بخش قبل ذکر شد، نتایج ساخت شبکه تحلیل استنادی در زبان فارسی با استفاده از روش ارائه شده قابل قبول و مناسب است. با این حال، همچنان مسائل و ایده‌هایی وجود دارند که در نظر گرفتن آن‌ها می‌تواند منجر به بهبود کیفیت روش پیشنهادی شود.

افزودن قابلیت استخراج خودکار فراداده‌های متناظر با پارساهای ورودی (شامل نام نویسنده، عنوان، دانشگاه یا پژوهشگاه محل انجام پژوهش و سال انجام پژوهش) از تمام متن آن‌ها مهم‌ترین مساله موجود در این زمینه است. در حال حاضر فرض بر این است که این داده‌ها در کنار تمام متن وجود دارند.

نکته دیگری که درباره روش پیشنهادی وجود دارد این است که در آن، فراداده‌های متناظر با رشته‌های مرجع در اولین مشاهده آن‌ها به دست آمده و در جدول اسناد ذخیره می‌شوند. با این حال، در بسیاری از موارد ارجاعات کامل نبوده یا تمام اقلام اطلاعاتی یک رشته مرجع به درستی توسط روش پیشنهادی شناسایی نمی‌شود. با توجه به این مساله، ممکن است اقلام اطلاعاتی به دست آمده در مشاهدات بعدی از آن رشته مرجع کامل‌تر باشند اما روش پیشنهادی این اقلام را در نظر نمی‌گیرد. به نظر می‌رسد که اصلاح جدول اسناد با توجه به مشاهدات بعدی رشته‌های مرجع می‌تواند منجر به بهبود روش پیشنهادی شود اما با توجه به گسترده بودن بحث، بررسی این مساله در این پژوهش صورت نگرفته و به پژوهش‌های آتی موکول شده است.

در پایان، می‌توان استفاده از دیگر دسته‌بندها مانند ماشین بردار پشتیبان دوقلو، CRF، یادگیری عمیق و همچنین دسته‌بندهای ترکیبی را برای حل مسائل این پژوهش را با هدف بهبود روش پیشنهادی مورد استفاده قرار داد.