به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی

عنوان طرح پژوهشی

طراحی یک روش هوشمند تجزیه رشته‌های مرجع در زبان فارسی

تاریخ شروع طرح پژوهشی	۱۳۹۶/۱۰/۰۴	**تاریخ پایان طرح پژوهشی**	۱۳۹۷/۱۱/۱۴
نام و نام خانوادگی مجری طرح پژوهشی	نصراله پاک‌نیت		
نام و نام خانوادگی همکاران طرح پژوهشی	جلال‌الدین نصیری، عمار جلالی‌منش		

دربارهٔ طرح پژوهشی

یک رشته مرجع را می‌توان به عنوان مجموعه‌ای از مولفه‌ها مانند نام نویسندگان، عنوان، محل نشر، سال نشر، شماره صفحات و ... در نظر گرفت. در حالیکه تجزیه رشته‌های مرجع موجود در انتهای یک مدرک علمی توسط کاربر انسانی به راحتی انجام‌پذیر است، تنوع موجود در فرمت‌های نگارش رشته‌های مرجع در کنار اشتباهات رخ داده توسط نویسندگان در نگارش این رشته‌ها، خودکارسازی انجام این عملیات را سخت کرده است. روش‌های زیادی برای خودکارسازی تجزیه رشته‌های مرجع ارائه شده است اما این روش‌ها وابسته به زبان بوده و امکان استفاده از یک روش ارائه شده برای یک زبان در زبانی دیگر منجر به نتایجی اشتباه می‌شود. تحقیقات صورت گرفته بیان‌گر این است که تاکنون هیچ روشی برای خودکارسازی تجزیه رشته‌های مرجع در زبان فارسی ارائه نشده است. با توجه به این مهم و نقش گسترده این مساله در ساخت خودکار شبکه‌های استنادی مدارک علمی و فرایندهای بازیابی اطلاعات، در این طرح پژوهشی به این مساله پرداخته شده است. در این راستا، پس از بررسی پیشینه مساله، از روش یادگیری ماشین بردار پشتیبان به عنوان یک دسته‌بند چند دسته‌ای استفاده شده و یک روش هوشمند برای مساله تجزیه رشته‌های مرجع در زبان فارسی ارائه شده است. با توجه به اهمیت انتخاب ویژگی‌های مناسب برای استفاده در دسته‌بند ماشین بردار پشتیبان، در این پژوهش، این مهم با توجه به ویژگی‌های استفاده شده در زبان انگلیسی و ویژگی‌های زبان فارسی و ارجاع‌دهی در این زبان انجام شده است. روش ارائه شده پیاده‌سازی شده و با استفاده از مجموعه داده‌ای شامل ۹۲۲ رشته مرجع فارسی ایجاد شده در این پژوهش آموزش داده شده و مورد آزمایش قرار گرفته است. نتایج به دست آمده نشانگر مقدار ۹۲٪ برای پارامترهای دقت، فراخوانی و اف-۱ می‌باشد.

روش پژوهش

در این پژوهش، با مطالعه منابع علمی معتبر روشی هوشمند پیشنهاد شده که با استفاده از آن بتوان تجزیه رشته‌های مرجع در زبان فارسی را ممکن ساخت. برای جمع‌آوری اطلاعات پژوهش، از روش کتابخانه‌ای استفاده شده و نتایج به‌کارگیری روش پیشنهادی روی مجموعه داده‌ای طراحی شده بدین منظور مورد ارزیابی و تحلیل قرار گرفته است. برای تحلیل نتایج از شاخص‌های مختلفی مانند دقت، فراخوانی و معیار اف-۱ استفاده شده است. لازم به ذکر است که مجموعه داده طراحی شده در این پژوهش شامل ۹۲۲ رشته مرجع بوده، شامل ارجاع به کتاب، مقاله همایش، مقاله نشریه و پایان‌نامه و رساله دانشگاهی است، که به صورت دستی و توسط خبره تجزیه شده است.

پیاده‌سازی روش ارائه شده با استفاده از زبان برنامه‌نویسی پایتون انجام شده و در ادامه از کامپیوتری با پردازنده مرکزی Intel i3" "3.7 GHz حافظیه داخلی "4 GB" و سیستم عامل ۶۴ بیتی ویندوز ۷ آزمایشاتی جهت بررسی کیفیت الگوریتم پیشنهادی استفاده شده است. نتایج آزمایشات انجام شده روی کامپیوتر ذکر شده نشان‌دهنده این است که آموزش روش پیشنهادی با توجه به مجموعه داده طراحی شده به کمتر از یک دقیقه زمان نیاز داشته و هر رشته مرجع را در زمانی تقریبا برابر با یک ثانیه تجزیه می‌کند.

یافته‌های طرح پژوهشی

در این پژوهش، در ابتدا به بررسی کارهای صورت گرفته در زمینه تجزیه رشته‌های مرجع در زبان انگلیسی پرداخته شد. بررسی‌های انجام شده نشان داد که روش‌های ارائه شده برای تجزیه رشته‌های مرجع را می‌توان به دو دسته کلی ۱- روش‌های مبتنی بر قاعده و ۲- روش‌های مبتنی بر یادگیری ماشین تقسیم‌بندی کرد که دسته دوم نتایج بسیار بهتری را ارائه کرده است. علاوه‌براین، از میان روش‌های یادگیری ماشین موجود، روش‌های مدل مخفی مارکوف (HMM)، میدان‌های تصادفی شرطی (CRF) و بردار ماشین پشتیبان (SVM) برای حل این مساله مورد استفاده قرار گرفته که استفاده از SVM و CRF منجر به دستیابی به نتایج بهتری شده است .

پس از بررسی پیشینه مساله، در این پژوهش، روشی مبتنی بر SVM برای تجزیه رشته‌های مرجع در زبان فارسی ارائه شد. برای طراحی روش ارائه شده در ابتدا مجموعه داده‌ای از رشته‌های مرجع برچسب‌گذاری شده در زبان فارسی تهیه گردید. سپس، بردارهای ویژگی متناظر با رشته‌های مرجع موجود در مجموعه داده تهیه شده ایجاد و دسته‌بند SVM با استفاده از آن آموزش داده شد. بردار ویژگی در نظر گرفته شده در روش ارائه شده را تلفیقی از ویژگی‌های مورد استفاده در زبان انگلیسی، برخی ویژگی‌های خاص زبان فارسی و همچنین نحوه استناددهی در این زبان است. نتایج بررسی آزمایشات بر روی دسته‌بند به دست آمده نشان‌دهنده ۹۲٪ برای هر سه پارامتر دقت، فراخوانی و اف-۱ می‌باشد.

نتیجه‌گیری و پیشنهادها

در این پژوهش، پس از بررسی پیشینه مساله، روشی مبتنی بر SVM برای تجزیه رشته‌های مرجع در زبان فارسی ارائه شد. نتایج بررسی آزمایشات بر روی دسته‌بند به دست آمده نشان‌دهنده ۹۲٪ برای هر سه پارامتر دقت، فراخوانی و اف-۱ می‌باشد. از جمله کارهای قابل انجام برای ارتقا این پژوهش می‌توان به استفاده از دسته‌بندهای دیگر مانند ماشین بردار پشتیبان دوقلو، CRF و یادگیری عمیق برای حل این مساله و مقایسه نتایج با نتایج به دست آمده در این پژوهش اشاره کرد. علاوه‌براین، همانطور که در مقدمه این پژوهش نیز ذکر شد، مساله تجزیه رشته‌های مرجع مساله‌ای پایه‌ای برای بسیاری مسائل دیگر است که از آن جمله می‌توان به تطبیق رشته‌های مرجع در یک سطح بالاتر و ساخت خودکار شبکه‌های استنادی در حالت کلی اشاره کرد. با توجه به کیفیت مناسب روش

ارائه شده، می‌توان از آن به عنوان ابزاری پایه‌ای برای حل این مسائل استفاده کرد که انشاالله در طرح‌های آتی به آن پرداخته خواهد شد.