بنیاد ملی نخبگان

۷

طرح شهید احمدی روشن

پیشنهاده

دوره هفتم

باسمه تعالی

سامانه‌ی جستجوی سه‌تایی‌های دانشی مبتنی‌بر یادگیری عمیق خوانش ماشینی

استاد خبره: علی میرعرب

استادیار، مدیر گروه اشاعه اطلاعات و تبادل دانش، پژوهشکده مدیریت اطلاعات و مدارک اسلامی، پژوهشگاه علوم و فرهنگ اسلامی

۱ **محل فعالیت خبره: پژوهشگاه علوم و فرهنگ اسلامی**

پژوهشگاه علوم و فرهنگ اسلامی

۲ **لینک صفحه شخصی خبره:**

http://islamicdoc.isca.ac.ir/Portal/home/?person/66582/745178/214600/%D9%87%DB%8C%D8%A7%D8%AA-%D8%B9%D9%84%D9%85%DB%8C

...

۳ **لینک فیلم توضیح پروژه توسط خبره در سایت آپارات:**

https://aparat.com/v/YMsCz

۴ **شرح موضوع طرح پیشنهادی:**

حجم عظیم داده‌های علوم اسلامی و توسعه روز افزون آن مشکلات فنی متعدی را برای بازیابی آن‌ها به همراه دارد. یکی از مشکلات رایج در فرآیند جستجوی گزاره، گزاره‌هایی هستند که با انواع حالات قابل بیان می‌باشند. گزاره‌های اتمی یا گزاره‌های سه‌قسمتی (موضوع، رابطه، محمول) از ساده‌ترین انواع این گزاره‌ها می‌باشد. مشکل ذکر شده حتی در جستجوی گزاره‌های اتمی (در عین سادگی) نیز وجود دارد. در نتیجه، یافتن دقیق اطلاعات مورد نیاز بدون درک نحو و معنایی محتوا دشوار است. رویکردهای متعددی برای حل این مشکل با استفاده از وب معنایی و تکنیک‌های داده‌های پیوندی پیشنهاد شده‌اند. در سال‌های اخیر پژوهش‌هایی در خصوص پیونددهی موجودیت و همچنین استخراج رابطه در علوم اسلامی صورت گرفته و پیکره‌های اولیه‌ای توسعه یافته و همچنین سامانه‌های یادگیرنده‌ی عمیقی نیز برای پیونددهی و استخراج رابطه توسعه داده شده است. همچنین، در سال‌های اخیر پژوهش‌های مهمی نیز در راستای جستجوی معنایی روی متون اسلامی صورت پذیرفته است. در این طرح، یک رویکرد مبتنی بر یادگیری عمیق برای جستجوی سه‌تایی‌های دانشی ارائه خواهیم کرد. در رویکرد ارائه شده ابتدا پیکره‌های اولیه توسعه یافته را پیش پردازش می کنیم تا آنها را به فرمت سه‌تایی تبدیل کنیم. دوم، تکنیک‌های مدل‌سازی ویژگی - کیسه کلمات (Bag of Words) و word2vec - برای یک یادگیری عمیق جدید

از پیکره ترکیب می‌شوند. سازوکار رویکرد این امکان را فراهم می‌کند تا تا معنای بین سه‌تایی‌ها درک شوند. سوم، یک شبکه عصبی کانولوشن برای بازیابی دقیق سه‌تایی‌ها با استفاده از یادگیری عمیق آموزش می‌دهیم. سرانجام یک موتور جستجو که قابلیت جستجوی گزاره‌های معنایی را (در عرض گزاره‌های لفظی) داشته باشد توسعه داده خواهد شد.

۵ هدف گیری خاص این طرح:

هدف اصلی این طرح توسعه سامانه موتور جستجوی معنایی با به‌کارگیری توأمان پیونددهی موجودیت و استخراج رابطه و برچسب‌نگاری معنایی پیکره متنی مورد جستجو با ابزارهای یادگیرنده عمیق است تا بتواند گزاره‌ها را مستقل از لفظ واژه(گان) و انواع حالات آنها جستجو کند. واسطی برای این سامانه طراحی خواهد شد که پژوهشگر بتواند گزاره‌هایی به زبان طبیعی مطرح کند و پس از تجزیه و تحلیل گزاره، نتایج دقیق و صحیح بازیابی شود.

۶ اهمیت انجام این طرح برای کشور:

در حال حاضر موسسات حوزوی و پژوهشی *متعددی* در کشور روی مفاهیم مختلف و روابط میان آنها در حال مطالعه هستند و یکی از دغدغه‌های اصلی آنها این است که روابط میان مفاهیم را در همه حوزه‌های پژوهشی اعم از اجتماعی، تربیتی، مذهبی، سیاسی و ... در اختیار داشته باشند و بتوانند به صورت جامع و دقیق در آن‌ها جستجو کنند تا بتوانند قضاوت و برداشت درست و صحیح از آنها ارائه کنند. مثلا رابطه میان مفهوم انسان و آزادی در هر یک از حوزه‌های یاد شده قابل بحث و بررسی است و هر یک از آنها متناسب با نگاه و بینش خود به آن می‌نگرند. چنانچه دید جامعی در تمامی این حوزه‌ها در اختیار پژوهشگر باشد، می‌تواند دقیق‌تر و با قضاوت درست‌تر به تحلیل این روابط بپردازد.

۷ کارفرما/مشتریان احتمالی پروژه:

کارفرما: مرکز مدیریت حوزه‌های علمیه

مشتری‌های احتمالی:

- ✓ عموم طلاب و پژوهشگران
- ✓ مؤسسات و مراکز حوزوی و دانشگاهی
- ✓ شرکت های فعال در حوزه هوش مصنوعی و داده کاوی

۸ کارهای مشابه انجام شده در داخل یا خارج کشور:

در (Soliman, 2020) روشی برای جستجو مبتنی بر یادگیری عمیق در گراف های RDF ارائه شده است. در این پژوهش درخواست‌های گراف‌های rdf به عنوان یک مشکل طبقه‌بندی در نظر گرفته شده است. رویکرد پیشنهادی یک طبقه‌بندی کننده یادگیری عمیق را بر روی مجموعه داده برای پیش‌بینی بازیابی نمودارهای RDF اعمال می‌کند. تران و همکاران

(Tran T, Wang H, Rudolph S, Cimiano P., 2009) ایده تولید نمودارهای خلاصه برای گراف RDF را برای تولید و رتبه بندی پرسش های SPARQL معرفی کرد. سپس، ژانگ و همکاران (Zhang L, Tran T, Rettinger A.,2013) راه حلی برای این ایده ارائه کرد. علاوه بر این، یانگ و همکاران (Yang M, Ding B, Chaudhuri S, Chakrabarti K, 2014) الگوهای درختی را برای اتصال کلمات کلیدی مشخص شده توسط کاربران پیشنهاد کردند که در آن الگوهای درختی بر اساس اندازه ارتباط آنها مرتب شده اند. ژنگ و همکاران (Zheng W, Zou L, Peng W, Yan X, Song S, Zhao D.,2016) روشی را برای جستجوی معنایی الگوهای ساختاری معادل پیشنهاد کرد. در نهایت، ویرجیلو (De Virgilio R. 2012) یک پرس و جو مبتنی بر کلمه کلیدی RDF از طریق محاسبه تنسور پیشنهاد کرد و سپس آن را از طریق MapReduce به یک محیط توزیع شده گسترش داد.

ناگاراجان و همکاران (Nagarajan G, Minu RI., 2015) سیستم بازیابی اطلاعات معنایی چند مدلی مبتنی بر هستی شناسی را ارائه کرد. این سیستم مبتنی بر ایده یکپارچه سازی دانش و تصاویر دامنه است و اطلاعات چند وجهی مورد نیاز را با استفاده از یک مجموعه قوانین فازی بازیابی می کند. نان و همکاران (To ND, Reformat MZ, Yager RR. ,2018) رویکردی را پیشنهاد کرد که درجات برابری بین روابط (خصوصیات) تعریف شده توسط واژگان مختلف را تعیین می کند. برای یافتن فواصلی که سطوح پایین و بالای برابری خصوصیت را نشان می‌دهند، وقوع تطبیق جفت‌های سه‌گانه RDF را در نظر می‌گیرند. در نتیجه، نموداری از خواص مشابه به دست می‌آید که در آن قدرت مبتنی‌بر فاصله‌ی لبه‌ها نشان دهنده درجاتی از شباهت بین ویژگی ها است.

در زبان فارسی پایگاه دانش زبان فارسی به صورت عمومی و چند دامنه ای مشتمل بر ۵۰۰ هزار موجودیت و ۷ میلیون رابطه میان آن ها با عنوان فارس بیس ارائه می گردد که به صورت متن باز در دسترس است. منابع اطلاعاتی فارس بیس عبارتند از: اطلاعات ساخت یافته ویکی پدیا مانند جعبه های اطلاعاتی، جداول وب و همچنین اطلاعاتی که توسط ماژول استخراج گر رابطه از متن خام استخراج شده اند. موجودیت های گراف دانش در یک هستان شناسی برگرفته از دی بی پدیا و سفارشی شده برای فارس بیس، سازمان دهی شده است. به منظور پیوند جعبه های اطلاعاتی ویکی پدیا به هستان شناسی بیش از ۷۰۰۰ نگاشت میان الگوها و خصیصه های ویکی پدیا با هستان شناسی برقرار شده است. همچنین با روش های یادگیری ماشین و با نظارت خبرگان، قسمتی از هستان شناسی و تعدادی از موجودیت ها به فارس نت متصل شده اند. مدل داده ای گراف دانش فارسی بر اساس استاندارد وب معنایی و به صورت RDF پیاده سازی شده است بنابراین داده ها به صورت سه تایی در پایگاه دانش ذخیره شده و می توان از طریق زبان SPARQL پرس وجوهای معنایی را بیان نمود.

۹ نیازمندی‌های این پروژه:

الف) نیازمندی نیروی انسانی:

تخصص‌های مورد نیاز	کارشناسی	کارشناسی ارشد	دکتری	توضیحات
مهندس کامپیوتر و هوش مصنوعی	۲	۲	۲	
متخصص علوم اسلامی	۱	۲		

ب) نیازمندی مالی و تجهیزاتی:

با وجود نیروی انسانی فوق و سخت‌افزارهایی که در اختیار است نیازی به تجهیزات نمی‌باشد.

۱۰ چشم‌انداز طرح و امکان توسعه:

توسعه موتور جستجویی که قابلیت جستجوی گزاره‌های معنایی را (در عرض گزاره‌های لفظی) داشته باشد. این موتور با طرح پرس و جوی مورد نظر کاربران به زبان طبیعی استاندارد، موجودیت‌ها و روابط میان آنها را به نحو مقتضی ارائه می‌دهد. در این موتور جستجو مفاهیم پرس‌وجو درک می‌شود و بازیابی اطلاعات سه‌تایی‌های دانشی به صورت مفهومی انجام می‌شود. در صورت محقق شدن اهداف این طرح امکان ارائه سرویس‌های بازیابی معنایی به موتورهای جستجوی فارسی، سیستم‌های پرسش و پاسخ و بانک‌های اطلاعاتی وجود دارد. از سوی دیگر پژوهشگران علوم اسلامی اگر ابزار جستجو معنایی در منابع خود را در اختیار داشته باشند می‌توانند با سرعت، دقت و جامعیت بیشتری به پژوهش بپردازند. همچنین می‌توان با به کارگیری تنظیمات فراپارمتری نتایج جستجو را بهبود داد.

۱۱ زمان‌بندی اجرای طرح:

ماه/کار	۱	۲	۳	۴	۵	۶	۷	۸	۹
پیش پردازش سه‌تایی‌ها									
آموزش سه‌تایی‌ها									
آزمایش سه‌تایی‌ها									
طبقه‌بندی کننده CNN									
ارزیابی خروجی‌ها									
انجام اصلاحات									