به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی برمبنای داده‌های فارسی پایگاه اطلاعات علمی ایران (گنج)

تاریخ شروع طرح پژوهشی	۱۳۹۸/۱۱/۱۷	**تاریخ پایان طرح پژوهشی**	۱۴۰۰/۱۲/۳
نام و نام خانوادگی مجری طرح پژوهشی	مرضیه زرین بال		
نام و نام خانوادگی همکاران طرح پژوهشی	آزاده محبی		

دربارهٔ طرح پژوهشی

سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که با ارائه پیشنهادهای متناسب با نیاز کاربر به او در تصمیم‌گیری کمک می‌کنند. نیازهای کاربر با کمک روش‌های مختلفی نظیر بهره‌گیری از عبارت پرس‌وجوی کاربر، تحلیل سند منتخب کاربر و یا تحلیل رفتار کاربر و یا کاربران مشابه استخراج می‌شود. سیستم پیشنهاددهنده مبتنی بر دانش یا مبتنی بر محتوی یکی از انواع سیستم‌های پیشنهاددهنده است و برای توسعه آن خصوصا در سال‌های اخیر از رویکردهای نوین نظیر یادگیری ماشین و تشخیص الگو استفاده می‌شود. خوشه‌بندی متن یکی از روش‌های بسیار پرکاربرد در یادگیری ماشین است که در آن اسناد مشابه براساس ویژگی‌های مناسب و شباهت بین اسناد خوشه‌بندی می‌شوند. لذا استخراج و انتخاب این ویژگی‌ها برای اسناد و خوشه‌بندی اسناد از جمله چالش‌های اصلی محسوب می‌شوند. همچنین، اغلب سیستم‌های پیشنهاددهنده متن و الگوریتم‌های خوشه‌بندی برای زبان انگلیسی توسعه پیدا کرده‌اند و به راحتی برای سایر زبان‌ها، مانند زبان فارسی، قابل بکارگیری نیستند. بنابراین هدف اصلی در این طرح، توسعه و بکارگیری رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی به کاربر است.

روش پژوهش

هدف اصلی در این طرح، توسعه و بکارگیری رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی به کاربر بود. بدین منظور ابتدا ادبیات رویکردهای خوشه‌بندی متن در سیستم‌های پیشنهاددهنده و ویژگی‌های مورد استفاده در هریک از رویکردها بررسی شده و رویکرد پیشنهادی معرفی شد. سپس با استفاده از فراداده‌های پایان‌نامه‌ها و رساله‌های فارسی پایگاه گنج، شامل اطلاعات مربوط به ۳۰۴۷ پایان‌نامه یا رساله فارسی رویکرد پیشنهادی پیاده-سازی و ارزیابی شد.

بر اساس ادبیات مربوط به خوشه‌بندی داده‌های بزرگ، سه معیار شباهت مینکویسکی، کسینوسی و آنتروپی نسبی به عنوان معیارهای فاصله یا شباهت انتخاب شدند. با کمک این معیارهای شباهت سه الگوریتم خوشه‌بندی K-means، FCM و REFCM در نرم‌افزار Python پیاده‌سازی شده و نتایج بدست آمد.

یافته‌های طرح پژوهشی

تمامی الگوریتم‌ها نتایج خوبی را حاصل کردند و در حالت c=11 و با استفاده از FCM (فاصله کسینوسی) و REFCM نتایج بهتری نسبت به c=13 حاصل شد. از آنجا که در این حالت هزینه محاسبات به مراتب کمتر c=13 خواهد بود از تفسیر خوشه‌های بدست آمده در حالت c=11 به عنوان مفسر داده‌های مورد پردازش استفاده شد. برمبنای این تفسیر، پایان‌نامه یا رساله‌های حوزه‌های زیر در یک خوشه مشترک قرار دارند و درصورتی که کاربر جستجویی را در یکی از زیرحوزه‌های زیر انجام داده باشد، نتایج مرتبط را می‌توان براساس دسته‌بندی زیر به او نمایش داد:

- ریاضی، برق، رباتیک
- سازه، عمران
- پزشکی، بیوشیمی، ژنتیک، انرژی، محیط زیست
- هنر، معماری، نقاشی
- تاریخ، فقه، ادبیات، رسانه
- حقوق، فلسفه، معارف، منطق، روابط بین‌الملل
- روانشناسی، آموزش
- کسب و کار، بازاریابی، مدیریت، دانش سازمانی، فناوری اطلاعات
- مسائل اجتماعی، شهرنشینی، گردشگری

- اقتصاد، بیمه
- مواد آلی و پلیمر، نانو، نور، پلاسما

نتیجه‌گیری و پیشنهادها

سیستم‌های پیشنهاددهنده پیشنهادهایی متناسب با نیاز کاربر به او ارائه داده و به تصمیم‌گیری برای انتخاب موارد مورد نیاز یا مورد علاقه کاربر کمک می‌کنند. تحلیل سند منتخب کاربر و استفاده از ویژگی‌های خاص آن برای پیشنهاد اسناد مشابه یکی از روش‌ها است.

براین اساس و در ادامه این پژوهش می‌توان به موضوع ساخت و توسعه سیستم پیشنهاددهنده‌ای براساس سیستم‌های پیشنهاددهنده مبتنی بر محتوی پرداخت که در آنها، ویژگی‌های سند(های) منتخب کاربر فعلی بررسی شده و مواردی به او پیشنهاد شود که با این ویژگی‌ها همخوانی یا شباهت داشته باشند.