پیش‌بینی عملکرد نتایج پرس‌وجو به‌کمک روش‌های بدون نظارت

سیده فاطمه کریمی[1]، مریم خدابخش[*2]
کارشناس‌ارشد مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران[*1]
استادیار دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران[2]

چکیده

در سال‌های اخیر، استفاده از موتورهای جست‌وجو افزایش روزافزون داشته و نیاز به توسعهٔ روش‌های دقیق‌تر بازیابی و رتبه‌بندی اسناد بیشتر شده‌است؛ درنتیجه پیش‌بینی عملکرد موتورهای جست‌وجو، یکی از الزامات و چالش‌های بازیابی اطلاعات محسوب می‌شود. اگر بتوان عملکرد پرس‌وجوها را پیش از مرحلهٔ بازیابی یا بعد از آن تخمین زد، می‌توان اقدامات خاصی را برای بهبود بازیابی انجام داد. پیش‌بینی عملکرد پرس‌وجو بر تخمین دشواری برآوردن درخواست کاربر برای یک روش بازیابی خاص متمرکز است. این پژوهش، به بررسی عملکرد پرس‌وجو به‌کمک روش‌های پس از بازیابی می‌پردازد؛ در این راستا از روش‌های بدون نظارت استفاده می‌شود و به خوشه‌بندی و اندازه‌گیری معیارهای مختلف جهت ارزیابی عملکرد پاسخ‌دهی پرس‌وجوها می‌پردازیم؛ درنهایت کار خود را با روش‌های بدون نظارت موجود در ادبیات این حوزه مقایسه خواهیم کرد. نتایج نشان می‌دهد روش پیشنهادی پژوهش حاضر توانست ضریب اسپیرمن را در مجموعه دادهٔ TREC DL 2019 و DL-Hard به‌ترتیب ۰.۰۰۹ و ۰.۱۶۳ و در مجموعه‌دادهٔ TREC DL 2020 ضریب پیرسون را ۰.۰۳۷ نسبت به بهترین کار موجود افزایش دهد.

واژگان کلیدی: پیش‌بینی عملکرد پرس‌وجو، بازیابی اطلاعات، موتورهای جست‌وجو.

Unsupervised Methods for Predicting Query Performance

Seyedehfatemeh Karimi[1], Maryam Khodabakhsh[*2]
M.Sc in Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran[1]
Assistant Professor of Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran[*2]

Abstract

With the rapid increase in the use of search engines, the need for developing more effective information retrieval and ranking methods has become critical. One of the key challenges in information retrieval is predicting query performance, which involves estimating how well a search engine can fulfill a user's information need. Accurate prediction of query performance allows search engines to take adaptive actions, such as query reformulation or ranking adjustment, to enhance retrieval effectiveness. Query Performance Prediction (QPP) methods fall into two main categories: pre-retrieval prediction and post-retrieval prediction. Pre-retrieval predictors estimate query difficulty before the retrieval process, relying on linguistic and statistical query features rather than retrieved documents. In contrast, post-retrieval prediction methods assess query performance based on the ranking list and document collection, providing deeper insights into retrieval effectiveness. In this study, we propose a novel unsupervised post-retrieval QPP method that evaluates query performance by analyzing the clustering behavior of retrieved documents. Our method defines five new metrics—CC, DCIC, DCNIC, DCNICR, and CCR— to measure the distribution and coherence of retrieved documents. These metrics help assess query difficulty by capturing how documents group into clusters, identifying outlier documents that do not fit well into clusters, and evaluating the overall structure of retrieved results. By leveraging these metrics, our approach provides a more fine-grained estimation of query performance without requiring human-labeled data. To evaluate the effectiveness of the proposed method, we conduct experiments on three datasets: TREC DL 2019, TREC DL 2020, and DL-Hard. The results demonstrate that our approach improves Spearman's correlation coefficient by 0.009 and 0.163 on the TREC DL 2019 and DL-Hard datasets, respectively. Additionally, it increases Pearson's correlation coefficient by 0.037 on the TREC DL 2020 dataset compared to state-of-the-art unsupervised QPP methods. These

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۲/۸/۲۲ • تاریخ پذیرش: ۱۴۰۳/۱۲/۱۸ • تاریخ انتشار: ۱۴۰۴/۳/۲۸ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.22.1.3]

improvements indicate that clustering-based QPP methods can effectively capture query difficulty and retrieval quality without the need for external supervision.

Keywords: Query Performance Prediction, Information Retrieval, Search Engines.

۱-مقدمه

در سال‌های اخیر، مطالعهٔ روش‌های طراحی‌شده برای جست‌وجوی اطلاعات به امری ضروری تبدیل شده‌است. با افزایش روزافزون استفاده از موتورهای جست‌وجو، نیاز به توسعهٔ روش‌های دقیق‌تر برای بازیابی و رتبه‌بندی اسناد بیش‌ازپیش احساس می‌شود[۱]؛ در همین راستا، بازیابی موقت به بازیابی یک فهرست رتبه‌بندی‌شده از اسناد به‌منظور برآورده‌کردن نیاز کاربران متمرکز است[۱، ۲، ۳]. روش‌های مختلفی برای بررسی اثربخشی بازیابی اطلاعات وجود دارند که در این میان روش‌های عصبی[۴-۹] عملکرد بهتری نسبت به روش‌های سنتی‌تر و مدل‌های زبان احتمال پرس‌وجو[1] [۲، ۱۰، ۱۱] دارند؛ با وجود این ،پژوهش‌ها نشان داده‌اند این روش‌ها برای تمام انواع پرس‌وجوها عملکرد مشابه ندارند[۱و۱۲] ؛ بدین ترتیب، پیش‌بینی عملکرد پرس‌وجو به تخمین عملکرد سامانه‌های بازیابی برای پرس‌وجوهای معین می‌پردازد[۱، ۲، ۱۳].

روش‌های پیش‌بینی عملکرد پرس‌وجو، کیفیت فهرست اسناد رتبه‌بندی‌شده و نیازهای اطلاعاتی کاربر را محاسبه می‌کنند[۱، ۲، ۱۴، ۱۵]. در سامانه‌های بازیابی اطلاعات اگر کاربر پرس‌وجو را به‌درستی مطرح نکند، موتور جست‌وجو می‌تواند با روش‌هایی مانند تولید یا فرمول‌بندی دوبارهٔ پرس‌وجو آن را اصلاح کند؛ به‌طور خاص، پیش‌بینی عملکرد پرس‌وجو با تخمین عملکرد سامانهٔ بازیابی، برای تشخیص دشواربودن یک پرس‌وجو طراحی می‌شود[۱۵، ۱۶]؛ به‌عبارت دیگر پیش‌بینی عملکرد پرس‌وجو به‌عنوان پیش‌بینی عملکرد یک سامانهٔ بازیابی اطلاعات بدون قضاوت‌های انسان تعریف می‌شود[۱۷]. روش‌های پیش‌بینی عملکرد پرس‌وجو می‌توانند با اهدافی مانند فرمول‌بندی مجدد پرس‌وجو [۱، ۱۶] یا مسیریابی پرس‌وجوها به رتبه‌بندی‌های جدید [۱، ۱۷] استفاده شوند.

عملکرد بازیابی اطلاعات حتی برای یک پرس‌وجوی یکسان، می‌تواند در سامانه‌های مختلف متفاوت باشد[۱۵، ۱۹، ۲۰]. با توجه به پژوهش‌های موجود در این زمینه، کارهایی که به پیش‌بینی عملکرد پرس‌وجو می‌پردازند به دو دستهٔ اصلی پیش‌بینی پیش از بازیابی و پیش‌بینی پس از بازیابی تقسیم می‌شوند[۱، ۲، ۱۳، ۱۴، ۱۶، ۲۱]. روش‌های پیش‌بینی پس از بازیابی براساس ارتباط بین پرس‌وجوی کاربر و اطلاعات موجود در مجموعه سند به پیش‌بینی عملکرد پرس‌وجو می‌پردازند[۱، ۲، ۱۷، ۲۱، ۲۲]. پیش‌بینی‌کنندگان پیش از بازیابی، عملکرد یک پرس‌وجو را پیش از رسیدن به مرحلهٔ بازیابی تخمین می‌زنند؛ بنابراین این روش‌ها مستقل از فهرست رتبه‌بندی نتایج‌اند[۱۴]؛ درحالی‌که روش‌های پیش‌بینی پس از بازیابی، یک پرس‌وجو را پس از بازیابی تجزیه‌وتحلیل می‌کنند تا در مورد دشواری پرس‌وجو قضاوت کنند[۱، ۲، ۱۷، ۲۱، ۲۲]. روش‌های پیش‌بینی پس از بازیابی، پیش‌بینی‌های خود را بر اساس فهرست رتبه‌بندی‌شدهٔ نتایج انجام می‌دهند؛ درنتیجه اطلاعات بیشتری در اختیار پیش‌بینی‌کننده قرار می‌دهند و دست‌یابی به پیش‌بینی‌های دقیق را آسان‌تر می‌کنند[۱۴].

مشارکت پژوهش حاضر در این مطالعه عبارت است از:

- یک رویکرد جدید مبتنی بر روش‌های بدون نظارت به‌منظور بررسی عملکرد پرس‌وجوها ارائه می‌شود.
- پنج معیار جدید CC، DCIC، DCNIC، DCNICR و CCR تعریف می‌شود که برای بررسی عملکرد پرس‌وجوها مورد استفاده قرار می‌گیرد.

چهارچوب پیشنهادی پژوهش حاضر در برابر روش‌های بدون نظارت با استفاده از مجموعه‌داده‌های استاندارد پرکاربرد از جمله TREC DL 2019، TREC DL 2020 و DL-Hard ارزیابی می‌شود؛ بدین‌منظور در بخش بعد به بررسی پژوهش‌های انجام‌شده در زمینهٔ پیش‌بینی عملکرد پرس‌وجو پرداخته می‌شود؛ سپس در بخش سه راه‌کار پیشنهادی بیان شده و در بخش چهارم به بررسی و مقایسهٔ نتایج به‌دست‌آمده پرداخته می‌شود؛ درانتها در بخش پنج به بیان نتیجه‌گیری و کارهای آینده پرداخته خواهد شد.

۲- پیشینهٔ پژوهش

همانطورکه در بخش پیشین مطرح شد، پژوهش‌های موجود در حوزهٔ پیش‌بینی عملکرد پرس‌وجو به دو دستهٔ کلیِ پیش از بازیابی و پس از بازیابی تقسیم می‌شوند. هر کدام از این رویکردها، به ارزیابی عملکرد پرس‌وجوها بر اساس معیارهای مختلفی می‌پردازند. پیش‌بینی‌کنندگان پس از بازیابی به‌کمک رویکردهای بدون نظارت یا با نظارت عملکرد پرس‌وجوها را پیش‌بینی می‌کنند. با توجه به اینکه پژوهش جاری بر پیش‌بینی‌کنندگان پس از بازیابی متمرکز است، در ادامه به معرفی برخی کارهای انجام‌شده در این دسته پرداخته می‌شود.

خدابخش و همکاران [۱] به معرفی یک رویکرد با نظارت برای پیش‌بینی عملکرد پرس‌وجو در پژوهش خود می‌پردازند. این پژوهش جهت بازیابی موقت و پیش‌بینی عملکرد پرس‌وجو مبتنی بر یادگیری چندوظیفه‌ای است. آن‌ها مدل BERT را که برای یادگیری ویژگی‌های مشترک دو وظیفه تعریف شده‌است، آموزش می‌دهند. این رویکرد متکی بر تنظیم دقیق

[1] Query Performance Prediction

[DOI: 10.61186/jsdp.22.1.3]

مدل BERT است تا کیفیت فهرست رتبه‌بندی‌شدهٔ اسناد بازیابی‌شده را برای پرس‌وجوی ورودی پیش‌بینی کند و سپس به رتبه‌بندی مجدد آن فهرست بپردازد.

عرب‌زاده و همکاران [۲] نیز یک رویکرد بانظارت جهت پیش‌بینی عملکرد پرس‌وجو معرفی می‌کنند. در این پژوهش به آموزش یک مدل می‌پردازند تا عملکرد پرس‌وجو را بر اساس ارتباط بین پرس‌وجو و اسناد بازیابی‌شده تخمین بزنند. آن‌ها به بررسی و تحلیل نتایج حاصل از دو معماری شبکهٔ cross-encoder و شبکهٔ bi-encoder در پژوهش خود پرداخته‌اند.

کرونن و همکاران [۲۳] یک رویکرد بدون نظارت را بیان می‌کنند که با توجه به مجموعه اسناد، انسجام و وضوح نتایج را محاسبه می‌کند. آن‌ها به ارائهٔ روشی برای پیش‌بینی عملکرد پرس‌وجو از طریق محاسبهٔ آنتروپی نسبی بین مدل زبانی پرس‌وجو و مدل زبانی مجموعه اسناد می‌پردازند. در این مطالعه با استفاده از امتیاز شفافیت[1] میزان ابهام یک پرس‌وجو اندازه‌گیری می‌شود. در این رویکرد رابطهٔ مستقیم بین امتیاز شفافیت و انحراف محاسبه‌شده وجود دارد و هرچه وضوح بیشتر باشد، پرس‌وجو آسان‌تر خواهد بود.

رایبر و همکاران [۲۴] رویکرد بازخورد پرس‌وجو[2] را معرفی کردند که در دستهٔ پیش‌بینی‌کنندگان بدون نظارت قرار می‌گیرد. در این پژوهش مدل کانال ارتباطی برای بازیابی اطلاعات در نظر گرفته می‌شود که از سه بخش پرس‌وجو، سامانهٔ بازیابی و فهرست نتایج تشکیل می‌شود. در این رویکرد، برای محاسبهٔ میزان استحکام، با استفاده از فهرست نتایج، مدل زبانی پرس‌وجو محاسبه می‌شود؛ سپس یک پرس‌وجوی جدید که به‌کمک واژگان با بیشترین احتمال در مدل زبانی ایجاد به سامانهٔ بازیابی داده می‌شود و نتایج جدید استخراج می‌شود. به‌کمک دو فهرست نتایج به‌دست‌آمده میزان استحکام محاسبه می‌شود و هر چه هم‌پوشانی اسناد بیشتر باشد، استحکام نیز بیشتر و پرس‌وجو ساده‌تر است.

رویتمن و همکاران [۲۵] یک پیش‌بینی‌کنندهٔ بدون نظارت را به نام افزایش اطلاعات وزن‌دار معرفی کردند که براساس امتیازهای نتایج به پیش‌بینی عملکرد پرس‌وجو می‌پردازد. رویکرد آن‌ها از یک روش امتیازدهی استفاده می‌کند و با توجه به پرس‌وجو به اندازه‌گیری اطلاعات مرتبط در اسناد مربوط به آن پرس‌وجو و اطلاعات مرتبط درکل اسناد می‌پردازد. این روش با استفاده از وزن‌دهی مناسب، امتیاز هر سند را به شکلی تنظیم می‌کند که کیفیت بازیابی آن را بازتاب دهد. نتایج ارزیابی‌ها نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های پیشین، عملکرد بهتری در پیش‌بینی عملکرد پرس‌وجو دارد؛ همچنین رویتمن و همکاران [۲۶] در پژوهش دیگری به ارائهٔ یک تخمین‌زنندهٔ جدید انحراف معیار برای پیش‌بینی عملکرد پرس‌وجو پس از بازیابی اطلاعات می‌پردازند. این تخمین‌زننده با استفاده از یک رویکرد نمونه‌برداری جدید توسعه یافته‌است که رفتار جست‌وجوی کاربر را شبیه‌سازی می‌کند. آن‌ها نشان می‌دهند که انحراف معیار نمرات بازیابی به‌عنوان یک شاخص قوی برای ارزیابی عملکرد پرس‌وجو عمل می‌کند. آن‌ها از چندین فهرست نتایج مرجع برای بهبود دقت پیش‌بینی عملکرد استفاده می‌کنند. نتایج نشان می‌دهد که روش پیشنهادی آن‌ها به‌طور قابل توجهی پیش‌بینی عملکرد پرس‌وجو را بهبود می‌بخشد.

در این بخش برخی کارهای انجام‌شده در حوزهٔ پیش‌بینی عملکرد پرس‌وجو بررسی شد. پژوهش‌های مختلف به‌منظور بررسی عملکرد پرس‌وجوها از روش‌های بانظارت یا بدون نظارت در کار خود استفاده کرده‌اند. استفاده از روش‌های بانظارت با چالش‌هایی روبه‌رو است؛ از جملهٔ این چالش‌ها می‌توان به دسترسی به داده‌های برچسب‌دار اشاره کرد. تهیهٔ داده‌های برچسب‌دار با کیفیت، نیازمند افراد خبره‌ای است که به مؤلفه‌های زبانی آشنا باشند؛ بنابراین برچسب‌گذاری داده‌ها یک کار سخت همراه با صرف زمان و هزینه است. به‌منظور مقابله با این چالش‌ها، کاهش استفاده از نیروی انسانی و جلوگیری از تأثیر خطاهای انسانی در آموزش مدل‌ها، در این پژوهش از روش‌های بدون نظارت برای پیش‌بینی عملکرد پرس‌وجوها استفاده خواهد شد. در بخش بعد راه‌کار پیشنهادی پژوهش حاضر شرح داده، سپس راه‌کار این پژوهش با کارهای موجود مقایسه می‌شود.

۳- روش پیشنهادی

هدف این پژوهش، پیش‌بینی عملکرد پرس‌وجو پس از بازیابی اسناد است؛ بدین منظور در این پژوهش رویکردی به‌کمک روش‌های بدون نظارت معرفی می‌شود. در ادامه راه‌کار خود را با جزئیات بیشتری بیان خواهیم کرد.

در رویکرد پیشنهادی، برای بررسی عملکرد پرس‌وجوها ویژگی‌های مشترک اسناد استخراج می‌شوند. جهت استخراج این ویژگی‌های مشترک، از روش‌های نظارت‌نشده استفاده شد. معماری مربوط به راه‌کار پیشنهادی در شکل (۱) قابل مشاهده است. در این روش ابتدا اسناد برتر مربوط به هر پرس‌وجو را بازیابی کرده و با یک لایه BERT [۲۷] اسناد رمزگذاری سپس خوشه‌بندی و در انتها معیارهایی تعریف شده و با محاسبهٔ این معیارها کیفیت پیش‌بینی‌های هر پرس‌وجو اندازه‌گیری می‌شوند.

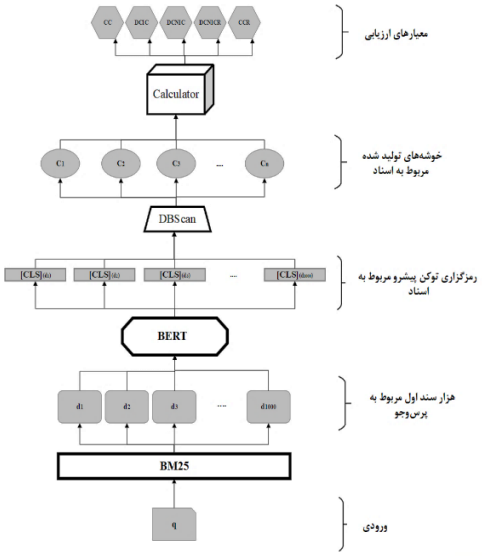
در نخستین مرحله، هزار سند برتر مربوط به هر پرس‌وجو بازیابی شدند. برای استخراج این اسناد از BM25 [۲۸] استفاده شد؛ سپس در مرحلهٔ بعد ویژگی‌های مشترک اسناد استخراج شدند؛ بدین منظور از یک لایه BERT استفاده و ارتباط هر سند به‌کمک توکن پیشرو[3] استخراج شد. با بررسی کارهای انجام‌شده، مدل‌های مبتنی بر ترنسفورماتورهای از پیش آموزش‌دیده مانند BERT [۲۷]،

[1] clarity score
[2] Query Feedback
[3] CLS

[DOI: 10.61186/jsdp.22.1.3]

دارای عملکرد خوبی در کار بازیابی موقت‌اند. از آنجا که این مدل‌های زبانی دارای توکن‌هایی هستند که می‌توانند معنای یک جمله را استخراج کنند؛ بنابراین این توکن‌ها برای کارهای رتبه‌بندی و دسته‌بندی اسناد می‌توانند استفاده شوند[۱].

برای سنجش سختی یک پرس‌وجو، کیفیت اسناد بازیابی‌شده ارزیابی می‌شوند؛ اگر اسناد مرتبط و هم‌موضوع باشند، پرس‌وجو ساده‌است، اما اگر اسناد موضوعات مختلفی داشته باشند، پرس‌وجو مبهم‌تر و پیچیده‌تر است؛ درنتیجه اگر پرس‌وجوها ساده و روشن باشند، نتایجی که برای آن‌ها بازیابی می‌شوند، همگی مرتبط به همان پرس‌وجو هستند؛ از طرفی دیگر اگر پرس‌وجو کلی و مبهم باشد، اسنادی که برای آن‌ها بازیابی می‌شوند به‌طورمعمول دارای موضوعات مختلف‌اند؛ درنتیجه اگر پس از خوشه‌بندی اسناد مربوط به هر پرس‌وجو، به تعداد خوشه‌های بیشتری دست پیدا کنیم به این معنی است که پرس‌وجوی مبهمی داریم و هر چه پرس‌وجو، کلی و مبهم باشد، سخت‌تر است. در پژوهش جاری از این ایده استفاده خواهیم کرد و برای سنجش عملکرد پرس‌وجوها به خوشه‌بندی اسناد بازیابی‌شده برای هر پرس‌وجو به‌منظور محاسبهٔ میزان سختی آن‌ها می‌پردازیم؛ بنابراین پس‌تولید رمزگذاری‌های مربوط اسناد به‌کمک توکن پیشروِ BERT، این رمزگذاری‌ها به‌عنوان ورودی جهت خوشه‌بندی استفاده می‌شوند. در این پژوهش از روش DBScan [۲۹] برای خوشه‌بندی اسناد استفاده شد؛ زیرا برخلاف روش‌هایی مانند K-Means، نیازی به تعیین تعداد خوشه‌ها از پیش ندارد؛ همچنین DBScan می‌تواند گروه‌هایی با اندازه‌های نامساوی را شناسایی کند، که برای تحلیل توزیع اسناد بازیابی‌شده مناسب‌تر است؛ بدین ترتیب اسناد مربوط به هر پرس‌وجو خوشه‌بندی و معیارهایی جهت رتبه‌بندی اسناد و بررسی عملکرد پرس‌وجوها محاسبه شدند. در ادامه به معرفی این معیارها خواهیم پرداخت.



(شکل-۱): معماری راه‌کار پیشنهادی در پژوهش جاری
(Figure-1): The Architecture of our Proposed Method.

در این پژوهش پنج معیار برای بررسی عملکرد پرس‌وجوها معرفی شده‌است. معیار CC@n به محاسبهٔ تعداد خوشه‌های مربوط به n سند نخست مربوط به هر پرس‌وجو می‌پردازد. تعداد بیشتر خوشه‌ها نشان‌دهندهٔ تنوع موضوعی در اسناد بازیابی‌شده است؛ بنابراین زمانی که تعداد خوشه‌ها زیاد باشد یعنی پرس‌وجوی مبهمی داریم که اسناد مختلف با موضوعات متنوع برای آن بازیابی شده‌است؛ درنتیجه هر چه CC بیشتر باشد نشان‌دهندهٔ مبهم‌بودن پرس‌وجو است.

$$CC@n = number\ of\ clusters\ of\ documents \quad (۱)$$

معیار DCIC@n تعداد اسناد خوشه‌بندی‌شده از میان n سند نخست را مشخص می‌کند. وقتی اسناد از نظر موضوع به پرس‌وجو شباهت داشته باشند در خوشه‌ها قرار می‌گیرند؛ درنتیجه هر چه تعداد داده‌های موجود در خوشه‌ها بیشتر باشد، یعنی سامانهٔ بازیابی به‌خوبی عمل کرده و اسناد مرتبط بازیابی شده‌است.

$$DCIC@n = number\ of\ documents\ in\ clusters \quad (۲)$$

معیار DCNIC@n از میان n سند نخست، تعداد اسنادی که در هیچ خوشه‌ای قرار نمی‌گیرند را مشخص می‌کند. زمانی که سندی در خوشه‌ها قرار نمی‌گیرد، یعنی از نظر موضوعی به پرس‌وجو شباهتی ندارد و به‌عنوان نوفه شناخته می‌شود؛ بنابراین مقدار کمتر DCNIC نشان‌دهندهٔ دقت بیشتر سامانهٔ بازیابی و ارتباط قوی‌تر اسناد بازیابی‌شده با پرس‌وجو است.

$$DCNIC@n = number\ of\ documents\ not\ in\ clusters \quad (۳)$$

همان‌طور که در فرمول (۴) قابل مشاهده است، معیار DCNICR@n نرخ اسناد خوشه‌بندی‌نشده به تعداد اسناد خوشه‌بندی‌شده، مربوط به n سند نخست هر پرس‌وجو را محاسبه می‌کند؛ هر چه این معیار کم‌تر باشد، یعنی تعداد اسناد خوشه‌بندی‌نشده کمتری داریم؛ بنابراین پرس‌وجو ساده‌تر و عملکرد سامانهٔ بازیابی بهتر است.

$$DCNICR@n = \frac{DCNIC@n}{DCIC@n} \quad (۴)$$

درنهایت معیار CCR@n نرخ خوشه‌های موجود مربوط به n سند نخست هر پرس‌وجو را محاسبه می‌کند؛ هر چه این معیار کمتر باشد؛ یعنی تعداد خوشه‌های کمتری داریم و اسنادی که خوشه‌بندی می‌شوند، در تعداد خوشه‌های کمتری قرار می‌گیرند؛ بنابراین پرس‌وجوی ساده‌تری داریم.

$$CCR@n = \frac{CC@n}{DCIC@n} \quad (۵)$$

در این بخش، به بررسی راه‌کار ارائه‌شده در پژوهش جاری پرداخته شد و معیارهایی برای بررسی عملکرد پرس‌وجوها معرفی شدند. در ادامه در نظر است راه‌کار پیشنهادی ارزیابی شود و نتایج کار با پژوهش‌هایی که به پیش‌بینی عملکرد پرس‌وجو به‌کمک روش‌های بدون نظارت می‌پردازند، مقایسه شود.

[DOI: 10.61186/jsdp.22.1.3]

۴- ارزیابی و تحلیل نتایج

در این بخش، ابتدا مجموعه‌داده‌های مورد استفاده در آزمایش‌های پژوهش معرفی می‌شوند؛ سپس روش‌های پایه معرفی خواهند شد. در آخر یافته‌ها به‌صورت کامل گزارش شده و تحلیل نتایج صورت می‌گیرد.

۴-۱-معرفی مجموعه‌داده

در این پژوهش، از سه مجموعه‌دادهٔ معروف در حوزهٔ پیش‌بینی عملکرد پرس‌وجو TREC DL 2019 [۳۰]، TREC DL 2020 [۳۱] و DL-Hard [۳۲] استفاده شده‌است. در جدول (۱) خلاصه‌ای از نتایج آماری این مجموعه‌داده‌ها گزارش شده‌است. مجموعه‌داده‌های TREC DL 2019، TREC DL 2020 و DL-Hard به‌ترتیب شامل ۴۳، ۵۴ و ۵۰ پرس‌وجو هستند که این مجموعه‌داده‌ها برای هر پرس‌وجو چندین سند مرتبط ارائه می‌دهند.

(جدول-۱): نتایج آماری مربوط به مجموعه داده‌های استفاده شده در پژوهش جاری

(Table-1): Statistics of the datasets used in current research.

	تعداد پرس‌وجو	تعداد ارتباط
TREC DL 2020	۵۴	۱۱۳۸۶
TREC DL 2019	۴۳	۹۲۶۰
DL-Hard	۵۰	۴۲۵۶

۴-۲-معرفی روش‌های پایه

همان‌طورکه پیش‌تر گفته شد، پژوهش جاری در حوزهٔ پیش‌بینی عملکرد پرس‌وجو پس از بازیابی قرار دارد؛ همچنین در این پژوهش از روش‌های بدون نظارت برای تشخیص کیفیت پرس‌وجوها استفاده می‌شود؛ درنتیجه به‌منظور ایجاد مقایسه‌ای منصفانه، کار خود را با بهترین روش‌های موجود که به پیش‌بینی عملکرد پرس‌وجو پس از بازیابی به‌کمک روش‌های بدون نظارت می‌پردازند، مقایسه می‌کنیم.

Clarity [۲۳]: یکی از رویکردهای پایه است که Divergence Leibler-Kullback بین مدل زبانی پرس‌وجو و مجموعه اسناد را محاسبه می‌کند.

WIG [۳۳]: در این رویکرد میزان اطلاعات مربوط به پرس‌وجو در اسناد را نسبت به میزان اطلاعاتی که به پرس‌وجو مرتبط است، محاسبه می‌کند.

HLDA: این روش که در [۳۴] ارائه شده‌است، یک روش تشخیص موضوع[1] است. از آنجا که هدف کار تشخیص موضوع با هدف کار این پژوهش شباهت زیادی داشت، از ایدهٔ تشخیص موضوعات استفاده شده‌است. با روش HLDA به تشخیص موضوعات اسناد پرداخته و از این طریق این اسناد خوشه‌بندی شدند.

QPP-PRP [۳۵]: این رویکرد به‌کمک روش‌های بدون نظارت، اطلاعاتی را که از طریق یک مدل رتبه‌بندی عصبی[2] استخراج می‌شود، به‌دست می‌آورد. ایدهٔ این پژوهش این است که هر چه احتمالات رتبه‌بندی یک سند نسبت به سند دیگر سازگارتر باشند، احتمال بهتربودن بازیابی و بهبود پیش‌بینی عملکرد پرس‌وجو بیشتر است.

۴-۳-معیارهای ارزیابی

در راه‌کار خود مشابه پژوهش [۱] از BM25 برای استخراج فهرستی از هزار متن نخست برای هر پرس‌وجو استفاده شد؛ بنابراین از معیارهای رسمی هر مجموعه‌داده جهت ارزیابی رویکرد بازیابی موقت استفاده می‌شود. معیار رسمی مربوط به سه مجموعه داده مورد بررسی NDCG@10 [۲۷، ۲۸] است؛ سپس برای ارزیابی پیش‌بینی عملکرد روش پیشنهادی خود، هم‌بستگی معیارهای معرفی‌شده در فرمول‌های (۱، ۲، ۳، ۴ و ۵) بین مجموعهٔ پرس‌وجوهایی که بر اساس عملکردشان مرتب شده‌اند و عملکرد واقعی آن‌ها که به وسیلهٔ NDCG به‌دست می‌آید، محاسبه شد؛ بدین ترتیب برای ارزیابی عملکرد روش پیشنهادی پژوهش ضرایب پیرسون[3]، کندال[4] و اسپیرمن[5] اندازه‌گیری خواهند شد. معرفی این ضرایب در ادامه می‌آید:

ضریب پیرسون یک متریک هم‌بستگی خطی است که نسبت بین کوواریانس دو فهرست و حاصل‌ضرب انحرافات استاندارد آن‌ها است:

$$Pearson = \frac{cov(M,\widehat{M})}{\sigma_M \, \sigma_{\widehat{M}}} \qquad (۶)$$

ضریب کندال، به‌عنوان یک متریک هم‌بستگی رتبه‌بندی، شباهت ترتیب‌بندی پرسش‌ها را زمانی که بر اساس عملکرد واقعی و پیش‌بینی‌شده رتبه‌بندی می‌شوند، اندازه‌گیری می‌کند:

$$Kendall = \frac{number\ of\ concordant\ pairs - number\ of\ discordant\ pairs}{number\ of\ pairs} \qquad (۷)$$

ضریب اسپیرمن یک معیار هم‌بستگی رتبه‌بندی است و به‌عنوان ضریب هم‌بستگی بین پرس‌وجوهای رتبه‌بندی‌شده تعریف می‌شود:

$$Spearman = \frac{cov((R(M),R(\widehat{M}))}{\sigma_{R(M)} \, \sigma_{R(\widehat{M})}} \qquad (۸)$$

[1] Topic Detection

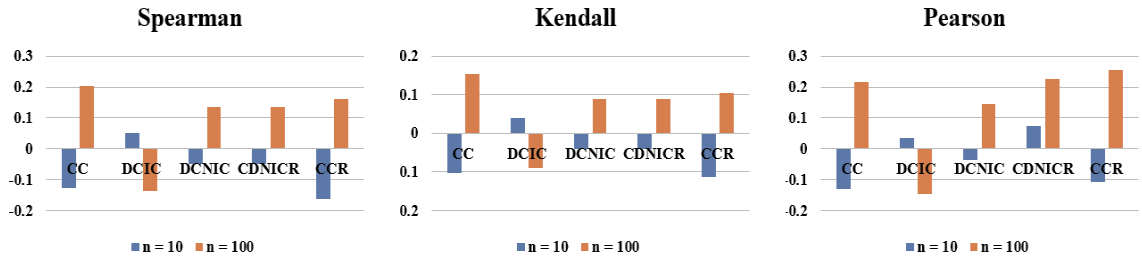
[2] neural ranking models
[3] Pearson
[4] Kendall
[5] Spearman

[DOI: 10.61186/jsdp.22.1.3]

(جدول-۲): عملکرد راه‌کار پیشنهادی برای ده و صد سند نخست بازیابی شده. اعتبار نتایج با t-test و سطح اطمینان ۹۵٪ تأیید شده‌است.

(Table-2): The performance of the proposed approach for the top 10 and 100 retrieved documents. The validity of the results has been confirmed using a two-tailed paired t-test at a 95% confidence level.

DL-Hard			TREC DL 2020			TREC DL 2019				
Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson		
-۰.۰۵۸	-۰.۰۴۲	-۰.۰۹۵	-۰.۱۲۷	-۰.۱۰۲	-۰.۱۲۹	۰.۰۹۸	۰.۰۷۵	۰.۱۳۲	CC	n =۱۰
-۰.۲۵۷	-۰.۱۹۳	-۰.۲۹۷	۰.۰۵	۰.۰۴	۰.۰۳۶	۰.۲۸۳	۰.۲۲	۰.۲۲	DCIC	
۰.۲۵۷	۰.۱۹۳	۰.۲۹۷	-۰.۰۵	-۰.۰۴	-۰.۰۳۶	-۰.۲۸۳	-۰.۲۲	-۰.۲۲	DCNIC	
۰.۲۵۷	۰.۱۹۳	۰.۲۸۹	-۰.۰۵	-۰.۰۴	۰.۰۷۴	-۰.۲۸۳	-۰.۲۲	-۰.۰۰۱	DCNICR	
۰.۱۳۸	۰.۰۹۸	۰.۰۹۹	-۰.۱۶۱	-۰.۱۱۳	-۰.۱۰۷	-۰.۲۰۲	-۰.۱۵۲	۰.۰۰۸	CCR	
۰.۰۳۶	۰.۰۲۳	۰.۰۲۵	۰.۲۰۳	۰.۱۵۳	۰.۲۱۶	۰.۲۳۶	۰.۱۸۷	۰.۰۰۴	CC	n =۱۰۰
-۰.۲۸۱	-۰.۱۹۹	-۰.۲۷۲	-۰.۱۳۷	-۰.۰۸۹	-۰.۱۴۶	۰.۲۲۲	۰.۱۵۲	۰.۲۶۳	DCIC	
۰.۲۸۱	۰.۱۹۹	۰.۲۷۲	۰.۱۳۷	۰.۰۸۹	۰.۱۴۶	-۰.۲۲۲	-۰.۱۵۲	-۰.۲۶۳	DCNIC	
۰.۲۸۱	۰.۱۹۹	۰.۲۱۵	۰.۱۳۷	۰.۰۸۹	۰.۲۲۶	-۰.۲۲۲	-۰.۱۵۲	-۰.۱۸۴	DCNICR	
۰.۱۱۹	۰.۰۷۸	۰.۱۱	۰.۱۶	۰.۱۰۴	۰.۲۲۵	۰.۰۷۵	۰.۰۶	-۰.۱۷۶	CCR	



(شکل-۱): مقایسهٔ ضرایب هم‌بستگی برای پنج معیار پیشنهادی در مجموعه‌دادهٔ TREC DL 2020 برای ده و صد سند نخست بازیابی‌شده.

(Figure-1): Comparison of correlation coefficients for the five proposed metrics on the TREC DL 2020 dataset for the top 10 and 100 retrieved documents.

(جدول-۳): مقایسهٔ نتایج روش پیشنهادی در پژوهش جاری با روش‌های بدون نظارت برای صد سند نخست بازیابی‌شده. اعتبار نتایج با t-test و سطح اطمینان ۹۵٪ تأیید شده‌است.

(Table-3): Comparison between the effectiveness of the proposed method in the current study and unsupervised methods for the top 100 retrieved documents. The validity of the results has been confirmed using a two-tailed paired t-test at a 95% confidence level.

DL-Hard			TREC DL 2020			TREC DL 2019				
Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson		
-۰.۰۴۶	-۰.۰۲۷	۰.۰۳۶	۰.۱۸۴-	-۰.۱۲۱	۰.۰۳۹	-۰.۰۶۸	-۰.۰۳۶	-۰.۲۹۶	Clarity	
۰.۱۱۸	۰.۱۶۲	۰.۱۱۳	۰.۱۳۳	۰.۰۸۶	۰.۱۱۷	۰.۱۶۱	۰.۱۰۷	۰.۱۵۱	WIG	
۰.۰۱۴	۰.۰۲۵	۰.۰۱۱	-۰.۰۷۲	-۰.۰۴۴	-۰.۰۹۴	-۰.۱۵۲	-۰.۰۹۱	-۰.۰۳۹	HLDA	
-۰.۰۰۶	۰.۰۱۸	-۰۰.۶۷	۰.۲۲۹	۰.۱۵۷	۰.۱۸۹	۰.۲۲۷	۰.۱۷۸	۰.۳۲۱	QPP-PRP	
۰.۰۳۶	۰.۰۲۳	۰.۰۲۵	۰.۲۰۳	۰.۱۵۳	۰.۲۱۶	۰.۲۳۶	۰.۱۸۷	۰.۰۴۴	CC	روش پیشنهادی
-۰.۲۸۱	-۰.۱۹۹	-۰.۲۷۲	-۰.۱۳۷	-۰.۰۸۹	-۰.۱۴۶	۰.۲۲۲	۰.۱۵۲	۰.۲۶۳	DCIC	
۰.۲۸۱	۰.۱۹۹	۰.۰۲۷۲	۰.۱۳۷	۰.۰۸۹	۰.۱۴۶	-۰.۲۲۲	-۰.۱۵۲	-۰.۲۶۳	DCNIC	
۰.۲۸۱	۰.۱۹۹	۰.۲۱۵	۰.۱۳۷	۰.۰۸۹	۰.۲۲۶	-۰.۲۲۲	-۰.۱۵۲	-۰.۱۸۴	DCNICR	
۰.۱۱۹	۰.۰۷۸	۰.۱۱	۰.۱۶	۰.۱۰۴	۰.۲۲۵	۰.۰۷۵	۰.۰۶	-۰.۱۷۶	CCR	

[DOI: 10.61186/jsdp.22.1.3]

۴-۴- نتایج ارزیابی

در این بخش به مقایسهٔ راه‌کار پیشنهادی پژوهش با روش‌های موجود در حوزهٔ پیش‌بینی عملکرد پرس‌وجو که از روش‌های بدون نظارت برای پیش‌بینی کیفیت پرس‌وجوها استفاده می‌کنند، پرداخته می‌شود.

در جدول (۲)، عملکرد معیارهای پنج‌گانهٔ تعریف‌شده در پژوهش جاری بر حسب ضرایب پیرسون، کندال و اسپیرمن برای مقادیر مختلف n جداگانه گزارش شده‌است. در سامانه‌های بازیابی اطلاعات و موتورهای جست‌وجو کاربران بر حسب معمول علاقه‌مند به بررسی اسناد نخست بازیابی‌شده‌اند؛ بنابراین از بررسی هزار سند بازیابی‌شده صرف نظر و نتایج کار برای ده و صد سند برتر مربوط به هر پرس‌وجو محاسبه شد؛ بهترین عملکرد به‌ازای مقادیر n در هر مجموعه‌داده به‌صورت پررنگ نشان داده شده‌است؛ همان‌طورکه در جدول (۲) نشان داده شده‌است، روش پیشنهادی برای صد سند نخست بازیابی‌شده عملکرد بهتری دارد و می‌تواند دست‌کم در یک معیار نتیجهٔ بهتری را ثبت کند. نتایج حاصل از هر دو بخش از نظر آماری بر اساس t-test با سطح اطمینان ۹۵٪ بررسی شدند و برای وظیفهٔ پیش‌بینی عملکرد پرس‌وجو معنادار است؛ علاوه‌براین، در شکل (۱) به مقایسهٔ نتایج معیارهای پنج‌گانه در مجموعه‌دادهٔ TREC DL 2020 پرداخته شد. نتایج نشان می‌دهند که عملکرد مدل پیشنهادی در پیش‌بینی کیفیت نتایج با افزایش تعداد اسناد بازیابی‌شده بهبود می‌یابد. در صد سند نخست، مقادیر هم‌بستگی‌های پیرسون، کندال و اسپیرمن در بیشتر موارد نسبت به ده سند نخست به‌طور قابل توجهی بهتر بوده‌است. این نتایج نشان‌دهندهٔ این است که بررسی تعداد بیشتری از اسناد می‌تواند به ارزیابی دقیق‌تر و بهینه‌تری از کیفیت نتایج بازیابی‌شده منجر شود؛ بنابراین در ادامه نتایج مربوط به صد سند نخست بازیابی‌شده با روش‌های پایهٔ بدون نظارت مقایسه می‌شوند.

جدول (۳) نتایج مربوط به راه‌کار ارائه‌شده در پژوهش جاری و روش‌های پایهٔ بدون نظارت را برای صد سند نخست بازیابی‌شده مقایسه می‌کند؛ درکل با توجه به نتایج به‌دست‌آمده، روش پیشنهادی حاضر توانست دست‌کم یک معیار از ضرایب پیرسون، کندال و یا اسپیرمن را با توجه به معیارهای پنج‌گانهٔ تعریف‌شده در پژوهش جاری در هر سه مجموعه‌داده بهبود ببخشد. راه‌کار پیشنهادی پژوهش حاضر توانست در تمامی مجموعه داده‌ها عملکرد بهتری نسبت به Clarity، WIG و HLDA به‌دست آورد.

با توجه به نتایج گزارش‌شده، روش ما در مجموعه‌دادهٔ DL-Hard عملکرد بهتری نسبت به روش QPP-PRP دارد. نتیجهٔ مربوط به ضریب پیرسون در مجموعه‌دادهٔ TREC DL 2019 و نتایج مربوط به ضرایب کندال و اسپیرمن در مجموعه‌دادهٔ TREC DL 2020 عملکرد بهتری داشته‌است. با مقایسهٔ نتایج بخش‌هایی که روش QPP-PRP را بهتر از راه‌کار پیشنهادی پژوهش حاضر عمل کرده‌است، مشاهده می‌شود که اختلاف نتایج گزارش‌شده بسیار کم است.

همان‌طورکه در جدول (۳) قابل مشاهده است، هر یک از معیارهای معرفی‌شده در این پژوهش شامل CC، DCIC، DCNIC، DCNICR و CCR براساس ویژگی‌های خاص هر مجموعه‌داده عملکرد متفاوتی دارند. در مجموعه‌دادهٔ TREC DL 2019، معیار CC عملکرد برجسته‌ای داشته‌است. این موضوع نشان می‌دهد که تنوع موضوعی اسناد بازیابی‌شده برای پرس‌وجوهای این مجموعه‌داده بیشتر است و خوشه‌بندی دقیق‌تر اسناد منجر به نتایج بهتری می‌شود. این امر به‌دلیل وجود پرس‌وجوهایی با پوشش موضوعی وسیع‌تر در این مجموعه‌داده است.

در مجموعه‌دادهٔ TREC DL 2020، معیارهای DCNICR و CCR به‌عنوان معیارهای کلیدی برجسته شده‌اند. این امر ناشی از توزیع متمرکزتر اسناد مرتبط در این مجموعه‌داده است که نیازمند ارزیابی دقیق‌تر نوفه و تعداد خوشه‌ها برای شناسایی پرس‌وجوهای ساده‌تر است.

در مجموعه‌دادهٔ DL-Hard، معیارهای DCNIC و DCNICR عملکرد بهتری داشته‌اند. این موضوع نشان می‌دهد که روش پیشنهادی حاضر توانسته است نوفهٔ اسناد بازیابی‌شده در پرس‌وجوهای سخت‌تر را بهتر مدیریت کند. این یافته‌ها اهمیت روش پیشنهادی در مواجهه با مجموعه‌داده‌هایی با پرس‌وجوهای چالش‌برانگیزتر را برجسته می‌کند.

روش پیشنهادی پژوهش حاضر در تمامی مجموعه‌داده‌ها توانسته است در حداقل یکی از معیارها عملکرد بهتری نسبت به روش‌های پایه ارائه دهد؛ برای مثال در مجموعه‌دادهٔ DL-Hard، معیار DCNIC توانسته است، نوفهٔ موجود در پرس‌وجوهای سخت را کاهش قابل توجهی دهد و اسناد مرتبط را با دقت بیشتری خوشه‌بندی کند. این ویژگی به‌دلیل طراحی دقیق معیارها و استفاده از خوشه‌بندی مبتنی بر ویژگی‌های BERT است که توانایی تمایز بین اسناد مرتبط و نامرتبط را بهبود می‌بخشد؛ درنتیجه انتخاب معیارهای مناسب برای ارزیابی باید بر اساس ویژگی‌های مجموعه‌داده و روش‌های پیشنهادی انجام شود؛ برای مثال، در پرس‌وجوهای سختِ مجموعه‌دادهٔ DL-Hard، معیارهای مرتبط با نوفه مانند DCNIC و DCNICR بیشترین اهمیت را دارند؛ درحالی‌که در پرسش‌های گسترده‌ترِ مجموعه‌دادهٔ TREC

[DOI: 10.61186/jsdp.22.1.3]

DL 2019، معیارهای مرتبط با تنوع موضوعی مانند CC اهمیت بیشتری پیدا می‌کنند.

۵- نتیجه‌گیری

در این پژوهش، روشی برای پیش‌بینی عملکرد پرس‌وجوها براساس خوشه‌بندی اسناد ارائه، سپس به مقایسه با روش‌های نظارت‌نشده پرداخته شد. با توجه به نتایج به‌دست‌آمده در این پژوهش، عملکرد خوشه‌بندی برای پیش‌بینی عملکرد پرس‌وجوها می‌تواند مؤثر باشد؛ همچنین با توجه به مجموعه‌دادهٔ مورد نظر، استفاده از معیارهای پنج‌گانهٔ تعریف‌شده می‌تواند مفید باشد. در نظر است در آینده با استفاده از روش‌های نظارت‌شده، نتایج بهبود یابد؛ همان‌طورکه پیش‌تر ذکر شد، در این پژوهش به معرفی پیش‌بینی‌کننده‌ای مبتنی بر روش‌های پس از ارزیابی پرداختیم؛ بااین‌حال، پیش‌بینی‌کنندگان پیش از ارزیابی نیز عملکرد خوبی در پیش‌بینی عملکرد پرس‌وجوها به‌دست آورده‌اند؛ درنتیجه نویسندگان در ادامهٔ در نظر دارند کار پژوهشی خود را با ترکیب روش‌های پیش از ارزیابی و پس از ارزیابی گسترش و بهبود بخشند.

۶-مراجع 6-References

[1] M. Khodabakhsh and E. Bagheri, "Learning to rank and predict: Multi-task learning for ad hoc retrieval and query performance prediction," *Inf. Sci. (Ny).*, vol. 639, 2023, doi: 10.1016/j.ins.2023.119015.

[2] N. Arabzadeh, M. Khodabakhsh, and E. Bagheri, "BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2857–2861, 2021, doi: 10.1145/3459637.3482063.

[3] J. ForutanRad, M. HourAli, and M. KeyvanRad, "Farsi Question and Answer Dataset (FarsiQuAD)," *Signal Data Process.*, vol. 20, no. 4, 2024, doi: 10.61186/jsdp.20.4.107.

[4] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *SIGIR 2020 - Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 39–48, 2020, doi: 10.1145/3397271.3401075.

[5] L. Xiong *et al.*, "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, pp. 1–16, 2021.

[6] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, *From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective*, vol. 1, no. 1. Association for Computing Machinery, 2022. doi: 10.1145/3477495.3531857.

[7] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020*, pp. 708–718, 2020, doi: 10.18653/v1/2020.findings-emnlp.63.

[8] T. A. Nakamura, P. H. Calais, D. de C. Reis, and A. P. Lemos, "An anatomy for neural search engines," *Inf. Sci. (Ny).*, vol. 480, pp. 339–353, 2019, doi: 10.1016/j.ins.2018.12.041.

[9] G. Hirst, J. Lin, R. Nogueira, and A. Yates, "Pretrained Transformers for Text Ranking: BERT and beyond," *Synth. Lect. Hum. Lang. Technol.*, vol. 14, no. 4, pp. 1–325, 2021, doi: 10.2200/S01123ED1V01Y202108HLT053.

[10] K. Sparck Jones, S. Walker, and S. E. Robertson, "Probabilistic model of information retrieval: Development and comparative experiments. Part 2," *Inf. Process. Manag.*, vol. 36, no. 6, pp. 809–840, 2000, doi: 10.1016/S0306-4573(00)00016-9.

[11] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," *SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval)*, vol. 51, no. 2, pp. 111–119, 2001, doi: 10.1145/383952.383970.

[12] E. Bagheri and F. N. Al-Obeidat, "A Latent Model for Ad Hoc Table Retrieval," *Adv. Inf. Retr.*, vol. 12036, pp. 86–93, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:215746638

[13] N. Arabzadeh, F. Zarrinkalam, J. Jovanovic, and E. Bagheri, "Neural embedding-based metrics for pre-retrieval query performance prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12036 LNCS, pp. 78–85, 2020, doi: 10.1007/978-3-030-45442-5_10.

[14] C. Hauff, D. Hiemstra, and F. De Jong, "A survey of pre-retrieval query performance predictors," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1419–1420, 2008, doi: 10.1145/1458082.1458311.

[15] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, and F. Scholer, "sMARE: a new paradigm to evaluate and understand query performance prediction methods," *Inf. Retr. J.*, vol. 25, no. 2, pp. 94–122, 2022, doi: 10.1007/s10791-022-09407-w.

[16] H. Roitman, S. Erera, and G. Feigenblat, "A study of query performance prediction for answer quality determination," *ICTIR 2019 - Proc. 2019 ACM SIGIR Int. Conf. Theory Inf. Retr.*, pp. 43–46, 2019, doi: 10.1145/3341981.3344219.

[17] G. Faggioli, T. Formal, S. Marchesin, S. Clinchant, N. Ferro, and B. Piwowarski, "Query Performance Prediction for Neural IR: Are We There Yet?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13980 LNCS, pp. 232–248, 2023, doi: 10.1007/978-3-031-28244-7_15.

[18] S. Sarnikar, Z. Zhang, and J. Zhao, "Query-Performance Prediction for Effective Query Routing in Domain-Specific Repositories," *J. Assoc. Inf. Sci. Technol.*, vol. 65, 2014, doi: 10.1002/asi.23072.

[19] D. Carmel and E. Yom-Tov, *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers, 2010.

[DOI: 10.61186/jsdp.22.1.3]

[20] J. S. Culpepper, G. Faggioli, N. Ferro, and O. Kurland, "Topic Difficulty: Collection and Query Formulation Effects," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, Sep. 2021, doi: 10.1145/3470563.
[21] G. Faggioli, S. Lupart, S. Marchesin, N. Ferro, and B. Piwowarski, "Towards Query Performance Prediction for Neural Information Retrieval : Challenges and Opportunities," pp. 51–63, doi: 10.1145/3578337.3605142.
[22] M. Khodabakhsh and E. Bagheri, "Semantics-enabled query performance prediction for ad hoc table retrieval," *Inf. Process. Manag.*, vol. 58, no. 1, p. 102399, 2021, doi: 10.1016/j.ipm.2020.102399.
[23] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," *SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval)*, pp. 299–306, 2002, doi: 10.1145/564426.564429.
[24] F. Raiber and O. Kurland, "Query-performance prediction: Setting the expectations straight," *SIGIR 2014 - Proc. 37th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 13–22, 2014, doi: 10.1145/2600428.2609581.
[25] H. Roitman, S. Erera, O. Sar-Shalom, and B. Weiner, "Enhanced mean retrieval score estimation for query performance prediction," *ICTIR 2017 - Proc. 2017 ACM SIGIR Int. Conf. Theory Inf. Retr.*, no. October, pp. 35–42, 2017, doi: 10.1145/3121050.3121051.
[26] H. Roitman, S. Erera, and B. Weiner, "Robust standard deviation estimation for query performance prediction," *ICTIR 2017 - Proc. 2017 ACM SIGIR Int. Conf. Theory Inf. Retr.*, pp. 245–248, 2017, doi: 10.1145/3121050.3121087.
[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
[28] M. and et al. Robertson, Stephen E and Walker, Steve and Jones, Susan and Hancock-Beaulieu, Micheline M and Gatford, "Okapi at TREC-3," *Nist Spec. Publ. Sp*, vol. 109, p. 109, 1995.
[29] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 169–194, 1998, doi: 10.1023/A:1009745219419.
[30] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos, "Overview of the TREC 2019 deep learning track," pp. 1–22, 2020, [Online]. Available: http://arxiv.org/abs/2102.07662
[31] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos, "Overview of the TREC 2020 deep learning track," pp. 1–13, 2021, [Online]. Available: http://arxiv.org/abs/2102.07662
[32] I. MacKie, J. Dalton, and A. Yates, "How Deep is your Learning: The DL-HARD Annotated Deep Learning Dataset," *SIGIR 2021 - Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. July, pp. 2335–2341, 2021, doi: 10.1145/3404835.3463262.
[33] Y. Z. and W. B. Croft, "Query performance prediction in web search environments," in *" in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 543–550.
[34] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," *Adv. Neural Inf. Process. Syst.*, no. May 2004, 2004.
[35] A. Singh, D. Ganguly, S. Datta, and C. Macdonald, *Unsupervised Query Performance Prediction for Neural Models with Pairwise Rank Preferences*, vol. 1, no. 1. Association for Computing Machinery, 2023. doi: 10.1145/3539618.3592082.

سیده فاطمه کریمی دانش‌آموختهٔ کارشناسی‌ارشد رشتهٔ مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه فردوسی مشهد و کارشناسی مهندسی کامپیوتر همین گرایش از دانشگاه صنعتی شاهرود است.
نشانی رایانامهٔ ایشان عبارت است از:
ftmkm9776@gmail.com

مریم خدابخش استادیار دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شاهرود هستند. ایشان دانش‌آموختهٔ مقطع دکترای تخصصی رشتهٔ مهندسی کامپیوتر، نرم‌افزار از دانشگاه فردوسی مشهد هستند. پیش از آن مدرک کارشناسی‌ارشد خود را در گرایش نرم‌افزار از دانشگاه فردوسی مشهد و مدرک کارشناسی خود را در همین گرایش از دانشگاه شهید بهشتی تهران دریافت کردند.
نشانی رایانامهٔ ایشان عبارت است از:
mkhodabakhsh@shahroodut.ac.ir

[DOI: 10.61186/jsdp.22.1.3]