مقاله پژوهشی

بهبود رتبه‌بندی با استفاده از BERT

شکوفه بستان، علی‌محمد زارع‌بیدکی و محمدرضا پژوهان

چکیده: رتبه‌بندی کارآمد اسناد در عصر اطلاعات امروز، نقش مهمی در سیستم‌های بازیابی اطلاعات ایفا می‌کند. این مقاله یک رویکرد جدید برای رتبه‌بندی اسناد با استفاده از مدل‌های درون‌سازی با تمرکز بر مدل زبانی BERT برای بهبود نتایج رتبه‌بندی ارائه می‌کند. رویکرد پیشنهادی از روش‌های درون‌سازی واژگان برای به‌تصویرکشیدن نمایش‌های معنایی پرس‌وجوهای کاربر و محتوای سند استفاده می‌کند. با تبدیل داده‌های متنی به بردارهای معنایی، ارتباط و شباهت بین پرس‌وجوها و اسناد تحت روابط رتبه‌بندی پیشنهادی با هزینه کمتر مورد ارزیابی قرار می‌گیرد. روابط رتبه‌بندی پیشنهادی عوامل مختلفی را برای بهبود دقت در نظر می‌گیرند که این عوامل شامل بردارهای درون‌سازی واژگان، مکان واژگان کلیدی و تأثیر واژگان باارزش در رتبه‌بندی بر مبنای بردارهای معنایی است. آزمایش‌ها و تحلیل‌های مقایسه‌ای برای ارزیابی اثربخشی روابط پیشنهادی اعمال گردیده است. نتایج تجربی، اثربخشی رویکرد پیشنهادی را با دستیابی به دقت بالاتر در مقایسه با روش‌های رتبه‌بندی رایج نشان می‌دهند. این نتایج بیانگر آن مسئله است که استفاده از مدل‌های درون‌سازی و ترکیب آن در روابط رتبه‌بندی پیشنهادی به‌طور قابل توجهی دقت رتبه‌بندی را تا ۰/۸۷ در بهترین حالت بهبود می‌بخشد. این بررسی به بهبود رتبه‌بندی اسناد کمک می‌کند و پتانسیل مدل درون‌سازی BERT را در بهبود عملکرد رتبه‌بندی نشان می‌دهد.

***کلیدواژه:* بردار معنایی، درون‌سازی واژه، رتبه‌بندی، یادگیری عمیق.**

۱- مقدمه

رتبه‌بندی اسناد بر اساس پرس‌وجوی کاربر، فرایندی است که در آن اسناد بر اساس میزان ارتباط با درخواست کاربر، رتبه‌بندی یا مرتب می‌شوند. هنگامی که کاربر یک پرس‌وجو را در یک موتور جستجو یا یک سیستم بازیابی اسناد وارد می‌کند، سیستم به تجزیه و تحلیل پرس‌وجو پرداخته و مجموعه‌ای از اسناد مرتبط را بازیابی می‌کند. با این حال، همه اسناد به یک میزان به پرس‌وجوی دریافتی مرتبط نیستند؛ بنابراین هدف از رتبه‌بندی، اولویت‌دهی به اسنادی است که به نیازهای اطلاعاتی کاربر نزدیک‌تر باشند. فرایند رتبه‌بندی معمولاً شامل عوامل متعددی است از جمله ارتباط محتوای سند با درخواست کاربر، اعتبار منبع، تازگی و محبوبیت. طی سال‌های اخیر، الگوریتم‌ها و روش‌های مختلفی برای ارزیابی ارتباط و رتبه‌بندی اسناد ارائه گردیده است. در شیوه سنتی، رتبه‌بندی اسناد بر اساس تطبیق واژه کلیدی است که در آن اسناد حاوی عبارات پرس‌وجو، رتبه بالاتری کسب می‌کنند. با این حال با پیشرفت در پردازش زبان طبیعی و یادگیری عمیق، رویکردهای جدیدتر با درک معنایی پرس‌وجو و محتوای سند ارائه گردیدند. در رویکرد درون‌سازی واژگان، هر واژه یا عبارت به‌صورت بردارهای متراکم نمایش داده می‌شود که می‌توان از آن در رتبه‌بندی بهتر اسناد استفاده نمود. به طور کلی، رتبه‌بندی اسناد بر اساس پرس‌وجوهای کاربر، جزئی حیاتی از سیستم‌های بازیابی اطلاعات است که تضمین می‌کند مرتبط‌ترین و مفیدترین اسناد بر اساس درخواست‌های جستجوی کاربران ارائه گردد.

مدل‌های سنتی اغلب بر اساس نمایش صریح ویژگی‌های متنی عمل می‌کنند؛ اما مدل‌های درون‌سازی از نمایش‌های توزیع‌شده و درون‌سازی واژگان، عبارات یا اسناد استفاده می‌کنند. مدل‌های درون‌سازی، این نمایش‌ها را از طریق روش‌های مبتنی بر شبکه عصبی مثل Word2vec، GloVe یا BERT یاد می‌گیرند و معنا و بافت معنایی کلمات و اسناد را در یک فضای برداری متراکم به تصویر می‌کشند. مدل‌های سنتی قادر به درک معنایی دقیق متن نیستند؛ زیرا در درجه اول بر ویژگی‌های سطحی و الگوهای آماری در متن تمرکز می‌کنند. اما مدل‌های درون‌سازی در درک معنایی متون و عبارات موفق‌تر عمل می‌کنند؛ زیرا با نگاشت واژگان یا اسناد به بردارهای متراکم می‌توانند روابط، شباهت‌ها و ارتباطات پیچیده معنایی را به تصویر بکشند. بنابراین در مدل‌های درون‌سازی از قدرت نمایش توزیع‌شده و الگوریتم‌های شبکه‌های عصبی استفاده می‌شود که در مقایسه با مدل‌های سنتی که بر ویژگی‌های صریح تکیه دارند، امکان درک معنایی و تعمیم پیشرفته‌تر را فراهم می‌کند.

تمرکز اصلی در رتبه‌بندی اسناد بر مبنای شباهت معنایی[1]، بر اولویت‌بخشیدن به اسنادی است که از نظر بافتی به پرس‌وجوی کاربر نزدیک‌تر هستند. در این رویکرد، میزان ارتباط دو واژه یا عبارت بر اساس معنی و بافت آن محاسبه می‌گردد. در این روش از مجموعه‌ای از ویژگی‌ها و معیارهای ازپیش‌تعریف‌شده مانند مترادف[2]، متضاد[3] و هم‌رخدادی[4] در یک مجموعه استفاده می‌شود. تشابه معنایی را می‌توان با استفاده از روش‌های مختلفی مانند معیارهای مبتنی بر محتوای اطلاعاتی و معیارهای توزیعی محاسبه نمود [۱]. رویکرد مبتنی بر مدل‌های درون‌سازی[5]، واژگان و عبارات را به‌عنوان بردار در فضایی با ابعاد بالا نشان می‌دهد تا فاصله بین بردارها بیانگر میزان شباهت معنایی آنها باشد. این مدل‌ها بر روی مقادیر زیادی از داده‌های متنی مانند مقالات خبری یا صفحات وب، آموزش داده می‌شوند تا روابط بین واژگان را بر اساس

این مقاله در تاریخ ۱۱ تیر ماه ۱۴۰۲ دریافت و در تاریخ ۱۰ آبان ماه ۱۴۰۲ بازنگری شد.

شکوفه بستان (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: sbostan@stu.yazd.ac.ir).

علی‌محمد زارع‌بیدکی، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: alizareh@yazd.ac.ir).

محمدرضا پژوهان، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: pajoohan@yazd.ac.ir).

1. Semantic Similarity
2. Synonyms
3. Antonyms
4. Co-occurrence
5. Embedding Model

الگوهای هم‌رخدادی بیاموزند. برای محاسبه شباهت میان واژگان و عبارات در مدل‌های درون‌سازی، اغلب از محاسبه شباهت کسینوسی بین دو بردار استفاده می‌شود [۲].

بسط پرس‌وجو[1] تکنیکی است که برای بهبود دقت مدل بازیابی استفاده می‌شود. از معایب این روش می‌توان به افزایش پیچیدگی محاسباتی[2] و کاهش دقت به‌دلیل بسط پرس‌وجو با واژگان نامرتبط یا بی‌ارزش اشاره کرد. در روش‌های مرسوم از درون‌سازی معنایی در گسترش پرس‌وجو استفاده می‌شود؛ اما در این مقاله از درون‌سازی معنایی متون به‌صورت مستقیم در رتبه‌بندی اسناد استفاده می‌گردد. در واقع به‌جای استفاده از نمایش برداری متون و استفاده از آن در گسترش پرس‌وجو، از بردارهای معنایی در فضای چندبُعدی و به‌صورت مستقیم در رتبه‌بندی استفاده می‌شود. به عبارت دیگر تمام محاسبات و سنجش میزان شباهت و ارتباط پرس‌وجو و اسناد، کاملاً مبتنی بر بردار معنایی متون در همان فضای چندبُعدی است. روش‌های درون‌سازی موجود، کاربردهای مختلفی همچون استفاده در گسترش پرس‌وجو دارند؛ اما اینکه یک پرس‌وجو یا عبارت به یک بردار درون‌سازی در فضای n بُعدی بدل شود و بر مبنای شباهتش با اسناد مورد بازیابی و رتبه‌بندی قرار گیرد، جای بحث دارد که در این پژوهش مورد بررسی، پیاده‌سازی و کاوش قرار می‌گیرد. به‌منظور دستیابی به بردار معنایی واژگان و متون، راهکارهای مختلفی مطرح گردیده که در ادامه به مرور آنها می‌پردازیم. سپس به معرفی روابط پیشنهادی جهت رتبه‌بندی بهتر اسناد بر مبنای پرس‌وجوی کاربر پرداخته می‌شود. روابط پیشنهادی، مبتنی بر بردارهای معنایی است و با تبدیل پرس‌وجو و سند به بردارهای با ابعاد بالا و بدون دخالت‌دادن سایر پارامترهای رایج در رتبه‌بندی، ضمن کاهش هزینه پردازشی به مرتب‌سازی اسناد بر مبنای درخواست کاربر و میزان ارتباط آن به پرس‌وجوی واردشده می‌پردازد.

ساختار مقاله به این ترتیب است که پژوهش‌های پیشین در بخش دوم بیان می‌گردد. بخش سوم به درون‌سازی با استفاده از مدل BERT سفارشی می‌پردازد. در بخش چهارم، روابط رتبه‌بندی پیشنهادی مطرح می‌گردند و در بخش پنجم، مجموعه دادگان مورد استفاده در رتبه‌بندی شامل جفت پرس‌وجو و اسناد با برچسب‌های میزان ارتباط هر سند به پرس‌وجوی مربوط تشریح می‌گردد. در بخش ششم، رتبه‌بندی با استفاده از درون‌سازی BERT مورد بررسی قرار می‌گیرد و نهایتاً در بخش هفتم، جمع‌بندی و نتیجه‌گیری نهایی مقاله بیان می‌گردد.

۲- پژوهش‌های پیشین

پژوهش‌های پیشین در قالب سه دسته بیان می‌گردند. در دسته اول پارامترهای رایج در رتبه‌بندی مطرح می‌گردد. دسته دوم به رویکرد درون‌سازی و انواع مدل‌های آن می‌پردازد. نهایتاً در دسته سوم، کارهای انجام‌شده در حوزه رتبه‌بندی اسناد بر مبنای بردارهای درون‌سازی مرور می‌گردد.

۲-۱ پارامترهای رتبه‌بندی

در رتبه‌بندی اسناد بر مبنای پرس‌وجوی کاربر، فرایند مرتب‌سازی نتایج جستجو بر اساس میزان ارتباط آن به پرس‌وجوی کاربر است. روش‌های متعددی برای رتبه‌بندی اسناد وجود دارد. در ادامه، برخی از عواملی که در محاسبه رتبه‌بندی اسناد به کار می‌روند بیان می‌گردد: ارتباط[3] که میزان مرتبط‌بودن سند به پرس‌وجوی کاربر را نشان می‌دهد. در رتبه‌بندی مبتنی بر ارتباط، نتایج جستجو به‌ترتیب نزولی بر مبنای میزان ارتباط سند به پرس‌وجوی کاربر مرتب می‌شوند. این ارتباط بر اساس عوامل مختلفی مانند تطابق[4] واژه کلیدی، مکان واژگان کلیدی، مترادف‌ها و پارامترهای دیگری همچون رفتار کاربر[5] تعیین می‌شود [۳].

محبوبیت[6] از پارامترهای مهمی است که بیانگر میزان محبوب‌بودن آن آن سند بر اساس میزان بازدیدها، لایک‌ها، نظرهای کاربران و زمان حضور کاربر[7] است. همچنین تازگی[8] سند که جدیدبودن سند بر مبنای تاریخ انتشار، فراوانی به‌روزرسانی‌ها و میزان تغییر در موضوع و بدنه سند را نشان می‌دهد، مورد استفاده قرار می‌گیرد. در رتبه‌بندی مبتنی بر رفتار کاربر، عوامل مختلفی همچون کلیک[9] و زمان حضور در صفحه در نظر گرفته می‌شود که اغلب بدین منظور از الگوریتم‌های یادگیری تقویتی[10] استفاده می‌گردد. همچنین در رتبه‌بندی مبتنی بر پروفایل کاربر[11]، علایق کاربر و سابقه مرور جستجو[12] در نظر گرفته می‌شود. سیستم‌های چندعاملی[13] اغلب برای جمع‌آوری داده‌های کاربر و پردازش آنها برای شناسایی علایق کاربر استفاده می‌شوند [۴]. در این مقاله تمرکز بر روی پارامتر ارتباط بین اسناد بوده و از سایر پارامترها استفاده نگردیده است.

۲-۲ انواع مدل‌های درون‌سازی

مدل‌های درون‌سازی، بازنمایی زبان را از متن خام یاد می‌گیرند که می‌تواند شکاف[14] بین واژگان پرس‌وجو و سند را پر کند [۵]. این مدل‌ها از طریق آموزش بر روی یک پیکره[15] بزرگ، بازنمایی معنایی عمیقی[16] را به‌دست می‌آورند که با بهره‌گیری از یادگیری انتقالی در طیف گسترده‌ای از وظایف پردازش زبان طبیعی[17] مانند شباهت اسناد[18]، خلاصه‌سازی متن[19]، طبقه‌بندی متن[20] و تحلیل احساسات[21] قابل استفاده است [۶]. بردارهای درون‌سازی می‌توانند با کسب دانش اضافی و بهبود نمایش واژگان و موجودیت‌ها در سند به رتبه‌بندی بهتر اسناد کمک کنند. برای تبدیل اسناد به بردارهای درون‌سازی معنایی و محاسبه شباهت بردارها، روش‌های مختلفی در زمینه پردازش زبان طبیعی وجود دارد. از روش‌های رایج می‌توان از درون‌سازی واژگان ازپیش‌آموزش‌یافته مانند الگوریتم‌های ایستای Word2vec، fastText یا Glove و همچنین الگوریتم‌های پویای درون‌سازی همچون ELMo، BERT یا GPT استفاده نمود.

1. Query Expansion
2. Computational Complexity
3. Relevance
4. Match
5. User Behavior
6. Popularity
7. Dwell Time
8. Freshness
9. Click
10. Reinforcement Learning Algorithms
11. User Profile Based Ranking
12. Search Browsing History
13. Multi Agent Systems
14. Gap
15. Corpus
16. Deep Semantic Representation
17. Natural Language Processing Tasks
18. Document Similarity
19. Text Summarization
20. Text Classification
21. Sentiment Analysis

جدول ۱: مقایسه مدل‌های درون‌سازی.

مدل	نقاط قوت	نقاط ضعف	ویژگی‌ها
Word2vec	– سرعت بالا – دقت بالا در پیداکردن روابط معنایی بین واژگان	عدم توانایی در درک معانی واژگان چندمعنایی و اصطلاحات	– مبتنی بر شباهت معنایی واژگان – استفاده از مدل CBOW و Skip-gram – محاسبه بردارهای واژگان بر اساس محتوای متن
fastText	– قابلیت کار با واژگان جدید – دقت بالا در پیداکردن روابط معنایی بین واژگان	– نیاز به داده‌های بزرگ برای آموزش – زمان طولانی برای آموزش – عدم توانایی در درک معانی واژگان چندمعنایی و اصطلاحات	– مبتنی بر شباهت معنایی واژگان – محاسبه بردارهای واژگان بر اساس محتوای متن و شکل واژگان
Glove	– دقت بالا در پیداکردن روابط معنایی بین واژگان – قابلیت کار با داده‌های کم‌حجم	– عدم توانایی در درک معانی چندمعنایی و اصطلاحات	– مبتنی بر شباهت معنایی واژگان – استفاده از ماتریس شباهت واژگان – محاسبه بردارهای واژگان بر اساس محتوای متن
ELMo	– قابلیت درک معانی چندمعنایی و اصطلاحات – دقت بالا در پیداکردن روابط معنایی بین واژگان – قابلیت کار با واژگان جدید	– نیاز به داده‌های بزرگ برای آموزش – زمان طولانی برای آموزش	– مبتنی بر شباهت معنایی واژگان و مدل‌های زبانی عمیق – استفاده از شبکه‌های عصبی بازگشتی – محاسبه بردارهای واژگان بر اساس محتوای متن و متن قبل و بعد از جمله
BERT	– قابلیت درک معانی چندمعنایی و اصطلاحات – قابلیت کار با واژگان جدید – دقت بالا در پیداکردن روابط معنایی بین واژگان – قابلیت استفاده در وظایف گوناگون مانند تشخیص احساسات و پرسش و پاسخ	– نیاز به داده‌های بزرگ برای آموزش – زمان طولانی برای آموزش	– مبتنی بر شباهت معنایی واژگان و مدل‌های زبانی عمیق – استفاده از شبکه‌های عصبی ترنسفورمر – محاسبه بردارهای واژگان بر اساس محتوای متن و متن قبل و بعد از جمله
GPT	– قابلیت تولید متن طبیعی – قابلیت کار با واژگان جدید – دقت بالا در پیداکردن روابط معنایی بین واژگان	– نیاز به داده‌های بزرگ برای آموزش – زمان طولانی برای آموزش	– مبتنی بر شباهت معنایی واژگان و مدل‌های زبانی عمیق – استفاده از شبکه‌های عصبی ترنسفورمر – محاسبه بردارهای واژگان بر اساس محتوای متن و متن قبل از جمله

الگوریتم GloVe یک مدل ازپیش‌آموزش‌یافته برای درون‌سازی واژگان است که از یک ماتریس هم‌زمانی برای تولید بردار هر واژه استفاده می‌کند [۷]. الگوریتم Word2vec در سال ۲۰۱۳ ارائه گردید که از یک مدل ازپیش‌آموزش‌یافته مبتنی بر شبکه عصبی استفاده می‌کند [۸]. در سال ۲۰۱۴، الگوریتم FastText [۹] توسط فیسبوک[1] مطرح گردید. در این الگوریتم از مدل Skip-gram الگوریتم Word2vec ایده گرفته شد؛ اما در آن از تابع وزن‌دهی متفاوتی استفاده گردیده است. در این روش، هر واژه به‌صورت کیفی از واژه‌ها به‌صورت n-gram در نظر گرفته می‌شود و از یک سری توکن[2] در آغاز و پایان هر واژه استفاده شده است. سپس به‌ازای تمام n-gramهای هر واژه، بردارهای عددی به شیوه مشابه با الگوریتم Word2vec به‌دست می‌آید و نهایتاً بردار هر واژه از مجموع تمامی بردارهای n-gram آن واژه حاصل می‌شود. مدل ELMo [۱۰] نوع جدیدی از نمایش واژه‌هاست که در سال ۲۰۱۸ معرفی گردید و به فهم عمیق معنایی و نحوی واژه‌ها می‌پردازد. برخلاف درون‌سازی‌های سنتی‌تری از واژه‌ها همچون Word2vec و GloVe، در مدل ELMo برای یک واژه نمایش‌های متفاوتی وجود دارد. در معماری ELMo از LSTM استفاده شده که یک نوع RNN است و به‌خوبی می‌تواند به‌عنوان یک مدل زبانی در نظر گرفته شود. دولین[3] و همکاران در سال ۲۰۱۸ الگوریتم مشهور BERT[4] را معرفی نمودند [۱۱] تا به بهبود دقت گوگل در کشف ساختار معنایی واژه‌های موجود در پرس‌وجوی کاربر کمک کنند. بزرگ‌ترین مدل زبان طبیعی منتشر شده در سال ۲۰۲۰ با نام GPT-۳ [۱۲] توسط OpenAI منتشر گردید که از لحاظ کارایی در ادامه GPT-۲ و GPT[5] است؛ با این تفاوت که تعداد پارامترها در آن به‌شدت افزایش یافته و روی داده‌های انبوه بسیار بزرگ‌تری نسبت به نسخه‌های قبلی آموزش دیده است. در این روش‌ها، واژه یا متن به بردار معنایی در فضایی با ابعاد بالا بر اساس الگوهای مشخص تبدیل می‌شود. همچنین می‌توان از مدل‌های یادگیری عمیق مانند حافظه کوتاه بلندمدت (LSTM)[6] برای درون‌سازی جملات استفاده نمود. این مدل‌ها بر روی مقادیر زیادی از داده‌های متنی، آموزش می‌یابند تا به کشف روابط بین واژگان و جملات دست یابند و درون‌سازی با کیفیت بالا ارائه دهند که می‌توان از آن برای محاسبه شباهت بین اسناد استفاده نمود [۱۳].

جدول ۱ به مقایسه مدل‌های مورد نظر و بیان نقاط قوت و ضعف مدل‌ها می‌پردازد. با توجه به جدول، مدل‌های Word2vec، fastText و Glove تنها برای محاسبه بردارهای واژگان بر اساس محتوای متن استفاده می‌شوند؛ در حالی که ELMo، BERT و GPT برای درک معنای واژگان در جمله و تولید متن نیز قابل استفاده هستند. مدل‌های fastText و ELMo قابلیت کار با واژگان جدید را دارند؛ در حالی که Word2vec و Glove این قابلیت را ندارند. مدل‌های ELMo، BERT و GPT نیاز به داده‌های بسیار بیشتری برای آموزش نسبت به سایر مدل‌ها دارند. همچنین مدل GPT قابلیت تولید متن دارد؛ در حالی که سایر مدل‌ها این قابلیت را ندارند. در این مقاله از مدل زبانی BERT به‌منظور درون‌سازی متون استفاده گردیده است.

1. Facebook
2. Token
3. Devlin
4. Bidirectional Encoder Representations from Transformers
5. Generative Pre-Training Transformer
6. Long Short-Term Memory

۲-۳ رتبه‌بندی اسناد بر مبنای بردارهای درون‌سازی

رتبه‌بندی اسناد بر مبنای مدل‌های درون‌سازی از کارهای مهم و جدید در سیستم‌های بازیابی اطلاعات است. پژوهش‌های فراوانی برای شناسایی بهترین مدل‌های درون‌سازی جملات و رتبه‌بندی بهتر آنها صورت پذیرفته است. در سال ۲۰۱۶، درون‌سازی موضوع مولد[1] ارائه گردید که ترکیبی از درون‌سازی واژه و مدل‌سازی موضوع[2] است. این مدل، اسناد را به‌عنوان بردارهای ویژگی با طول ثابت در یک فضای پیوسته کم‌بعد[3] تحت محور موضوع[4] نمایش می‌دهد؛ بنابراین احتمال هر واژه تحت تأثیر بافت محلی[5] و موضوع آن است [۱۴]. الگوریتم DESM[6] در سال ۲۰۱۶ ۲۰۱۶ به‌عنوان یک مدل فضای درون‌سازی دوگانه برای رتبه‌بندی اسناد مبتنی بر الگوریتم Word2vec ارائه گردید که به آموزش واژه‌ها در سند و پرس‌وجو می‌پردازد [۱۵]. دهقانی و همکاران در سال ۲۰۱۷ به ارائه یک مدل عصبی با نظارت ضعیف[7] پرداختند. در این روش از خروجی یک مدل مدل رتبه‌بندی بدون نظارت مانند BM۲۵ به‌عنوان یک سیگنال نظارت ضعیف استفاده گردید [۱۶]. K-NRM یک مدل عصبی مبتنی بر هسته برای رتبه‌بندی اسناد است که از یک ماتریس ترجمه[8] برای مدل‌سازی شباهت‌های سطح واژه[9] از طریق درون‌سازی واژگان استفاده می‌کند که در سال ۲۰۱۷ ارائه گردید [۱۷]. در سال ۲۰۱۹، یک رویکرد فاکتورسازی ماتریس[10] برای درون‌سازی نودها در شبکه‌ای از اسناد ارائه گردید. این رویکرد الهام‌گرفته از الگوریتم GloVe[11] است که مبتنی بر احتمال هم‌رخدادی واژه‌ها[12] است [۱۸]. در سال ۲۰۲۰ روشی به نام درون‌سازی گوسی از اسناد پیوندی (GELD)[13] معرفی گردید که به درون‌سازی اسناد پیوندی به یک فضای معنایی پیش‌آموزش‌یافته می‌پردازد که شامل مجموعه‌ای از بردارهای درون‌سازی‌شده است [۱۹]. در سال ۲۰۲۰ رویکرد نوینی با هدف بهبود رتبه‌بندی اسناد جستجو بر مبنای سنجش شباهت معنایی[14] و عامل ارتباط[15] ارائه گردید. شباهت معنایی بر بازیابی اسناد مشابه متنی بر اساس یک پرس‌وجو متمرکز است؛ در حالی که عامل ارتباط بر ساخت یک مدل عصبی مبتنی بر ادغام هسته[16] تمرکز دارد [۲۰]. در سال ۲۰۲۲، درون‌سازی موجودیت‌ها مبتنی بر ارتباط[17] به‌منظور رتبه‌بندی بهتر اسناد مورد بررسی قرار گرفت و از یک شبکه عصبی برای آموزش درون‌سازی اسناد ویکی‌پدیا بر پایه گراف استفاده گردید [۲۱]. یک الگوریتم یادگیری ماشین بر مبنای رتبه‌بندی مجدد اسناد در سال ۲۰۲۲ معرفی شد. ساختار رتبه‌بندی به این صورت است که در ابتدا پرس‌وجوها و اسناد با استفاده از الگوریتم BERT رمزگذاری می‌شوند و سپس از یک مدل یادگیری رتبه‌بندی (LTR)[18] مجدد اسناد مبتنی بر TFR[19] برای بهبود نتایج و بهینه‌سازی بیشتر عملکرد رتبه‌بندی استفاده می‌شود [۲۲].

به‌منظور رتبه‌بندی اسناد بر مبنای بردارهای درون‌سازی در این مقاله، روابط پیشنهادی در ادامه ارائه می‌گردد که در آن، پرس‌وجو و سند به بردارهای درون‌سازی تبدیل می‌گردند و در همان فضای n بُعدی بر اساس روابط پیشنهادی مورد سنجش قرار گرفته و اسناد با امتیاز بالاتر، اولویت‌دهی می‌شوند.

۳- درون‌سازی با استفاده از مدل BERT سفارشی

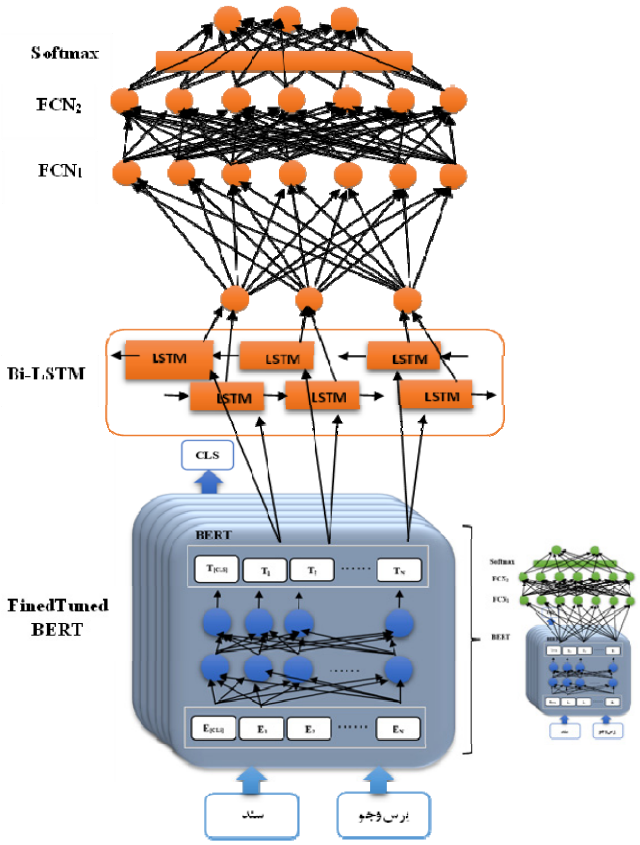
بستان و همکاران در سال ۲۰۲۳، یک مدل برت سفارشی ارائه دادند که به آموزش بردارهای درون‌سازی بر روی وب فارسی می‌پرداخت [۲۳]. این مدل شامل یک مرحله پیش‌آموزش مدل و دو مرحله تنظیم دقیق است. در این روش، سه مدل پیش‌آموزش‌یافته مورد استفاده قرار گرفت و مدل اول از طریق آموزش اولیه بر روی وب فارسی تهیه گردید. دو مدل دیگر از مدل‌های معروف برت هستند که قبلاً آموزش یافته و در دسترس عموم قرار گرفته‌اند. همچنین مدل برت سفارشی که بر روی صفحات وب فارسی به‌صورت سفارشی آموزش‌یافته است در این ارزیابی مورد استفاده قرار می‌گیرد. مدل برت چندزبانه که زبان فارسی را پوشش می‌دهد و مدل پارس برت [۲۴] که مبتنی بر زبان فارسی آموزش‌یافته است، از مدل‌های مورد استفاده در این ارزیابی هستند. فرایند تنظیم دقیق طی دو مرحله متوالی اعمال گردیده و سه مدل درون‌سازی‌شده تولید گردیدند.

در واقع در این ارزیابی، سه مدل استفاده می‌شوند که طی سه مرحله، آموزش یافته‌اند. مرحله اول پیش‌آموزش مدل است که تنها مدل برت سفارشی بر مبنای صفحات وب فارسی و به‌صورت سفارشی آموزش یافته است. دو مدل پیش‌آموزش‌یافته دیگر، قبلاً توسط ارائه‌دهندگان مدل برت و سایر محققان مورد آموزش قرار گرفته‌اند و قابل استفاده هستند. در این راستا مراحل تنظیم دقیق سه مدل پیش‌آموزش‌یافته طبق معماری مطرح در شکل ۱، توسط بستان و همکاران [۲۳] ارائه گردیده که به‌عنوان مدل‌های نهایی در این ارزیابی مورد استفاده قرار می‌گیرد.

٤- روابط رتبه‌بندی پیشنهادی

هدف از این پژوهش، رتبه‌بندی بهتر اسناد بر مبنای بردارهای معنایی با صرف هزینه کمتر و دقت بالاتر است. برای این کار از محاسبه شباهت بردارهای معنایی پرس‌وجو و اسناد استفاده می‌گردد. ایده این مقاله در استخراج بردارهای معنایی واژگان و استفاده از آن در فرمول‌های رتبه‌بندی با رویکرد جدید است. در این رویکرد، بردارهای معنایی که در فضای چندبُعدی ارائه شده، در همان فضا مورد رتبه‌بندی قرار می‌گیرند. این رتبه‌بندی بر مبنای محاسبه کسینوس زاویه بین دو بردار اما با بهره‌گیری از ساختار جدید و طی روابط پیشنهادی در فضای چندبُعدی اعمال می‌گردد. جهت استخراج بردارهای معنایی عبارات پرس‌وجو و سند در این مقاله از مدل‌های سفارشی فارسی آموزش‌یافته BERT استفاده می‌گردد [۲۳]. نکته قابل توجه در معماری BERT بر دریافت یک عبارت یا جمله و ارائه بردار معنایی آن عبارت است. در مدل BERT، یک جمله یا دنباله‌ای از واژگان به‌عنوان ورودی دریافت می‌شود و نمایش بافت آن جمله در قالب یک بردار درون‌سازی‌شده به‌عنوان خروجی تولید می‌گردد؛

1. Generative Topic Embedding
2. Topic Modeling
3. Low Dimensional Continuous Space
4. Topic
5. Local Context
6. Dual Embedding Space Model
7. Weak Supervision
8. Translation Matrix
9. Model Word Level Similarities
10. Matrix Factorization Approach
11. Global Vectors for Word Representation
12. Co-occurrence Probabilities of Words
13. Gaussian Embedding of Linked Documents
14. Semantic Similarity
15. Relevance Factor
16. Kernel Pooling
17. Relevance Based Entity Embedding
18. Learning to Rank Model
19. TensorFlow Ranking



شکل ۱: مدل BERT سفارشی طی فرایندهای پیش‌آموزش و تنظیم‌های دقیق متوالی [۲۳].

اما در این مقاله به‌جای استفاده از خروجی کل مدل که یک بردار درون‌سازی‌شده به‌ازای جمله ورودی است، بردارهای درون‌سازی هر واژه از جمله ورودی به‌صورت جداگانه و از طریق لایه‌های پنهان نهایی استخراج می‌گردد. بنابراین هر واژه در جمله ورودی، بازنمایی متنی یا بردار درون‌سازی خود را خواهد داشت. با بهره‌گیری از این ویژگی می‌توان از اطلاعات متنیِ غنی به‌دست‌آمده در لایه‌های پنهان مدل نیز بهره‌مند گردید. در واقع به‌منظور کنترل بیشتر روی مدل و کسب دقت بیشتر در رتبه‌بندی، به‌ازای هر واژه موجود در جمله ورودی، بردارهای معنایی آن از طریق استخراج لایه‌های پنهان قبل از لایه نهایی استخراج می‌گردد. برای دریافت بردارهای درون‌سازی واژگان یک جمله با استفاده از مدل BERT از نشانه‌ساز BERT استفاده می‌گردد. نشانه‌ساز، جمله ورودی را به لیستی از واژگان تبدیل می‌کند و سپس این لیست را به درون مدل BERT می‌فرستد تا دنباله‌ای از حالت‌های پنهان تولید گردد. نهایتاً بردارهای درون‌سازی واژگان از طریق حاصل‌ضرب نقطه‌ای حالت‌های پنهان با ماتریس وزن آموزش‌یافته مدل به‌دست می‌آیند. بنابراین از مدل BERT سفارشی به‌صورت متفاوت استفاده می‌گردد و خروجی سفارشی بر مبنای واژگان مد نظر استخراج می‌شود.

در صورت استفاده از لایه نهایی مدل و بدون درنظرگرفتن بردار درون‌سازی واژگان به‌صورت مجزا بر اساس جمله ورودی T که می‌تواند مطابق (۱) شامل سند d یا پرس‌وجوی q باشد، بردار درون‌سازی کل عبارت از روی مدل M_{BERT} مورد ارزیابی، مطابق (۲) استخراج می‌گردد

$$q, d \in T \qquad (۱)$$

$$\vec{e} = M_{BERT}(T) \qquad (۲)$$

جهت محاسبه شباهت میان بردار اسناد می‌توان از معیارهای سنجش شباهت مانند شباهت کسینوسی، فاصله اقلیدسی یا فاصله منهتن استفاده کرد. شباهت کسینوسی، معیاری محبوب برای اندازه‌گیری کسینوس زاویه بین دو بردار با رنج عددی بین منفی و مثبت یک است. هرچه مقدار به‌دست‌آمده به عدد یک نزدیک‌تر باشد، بیانگر شباهت بیشتر دو سند مربوط به یکدیگر است [۲۵]. بنابراین شباهت دو بردار معنایی طبق (۳) از طریق محاسبه کسینوس زاویه بین بردارها محاسبه می‌گردد. این شباهت به‌ازای هر پرس‌وجو و سند محاسبه می‌گردد و سپس اسناد بر مبنای شباهت به‌دست‌آمده، مرتب و رتبه‌بندی می‌گردند. رتبه‌بندی اسناد بر مبنای این روش، SentenceSim نامیده می‌شود

$$Similarity = \cos\theta = \frac{\vec{e}_q . \vec{e}_d}{|\vec{e}_q||\vec{e}_d|} \qquad (۳)$$

اما در ادامه، روابط رتبه‌بندی پیشنهادی با بهره‌گیری از بردارهای درون‌سازی هر واژه از جمله ورودی، از طریق لایه‌های پنهان نهایی استخراج می‌گردد تا در رتبه‌بندی بهتر اسناد مورد ارزیابی قرار بگیرد. هر رابطه در تکمیل رابطه قبلی و در راستای بهبود نتایج رتبه‌بندی گام می‌نهد.

۴-۱ رابطه DocCentroidSim

در این رابطه به رتبه‌بندی بر مبنای میانگین کسینوس هر واژه پرس‌وجو و نقطه مرکزی سند[1] پرداخته می‌شود. ابتدا در فاز برون‌خط به‌ازای هر سند، بردار صدبُعدی تک‌تک واژه‌های آن سند از مدل درون‌سازی به‌دست‌آمده استخراج شده و با محاسبه میانگین آنها، بردار جدیدی که بیانگر نقطه مرکزی آن سند است به‌دست می‌آید. بردار به‌دست‌آمده به‌ازای تمام اسناد موجود بر روی فایل ذخیره می‌گردد تا در زمان اجرای برخط مورد استفاده قرار گیرد. سپس به‌ازای هر واژه پرس‌وجو، بردار صدبُعدی آن از مدل مورد نظر استخراج و شباهت کسینوسی آن با بردار نقطه مرکزی سند، محاسبه می‌گردد و بعد از محاسبه به‌ازای تمام واژه‌های پرس‌وجو، میانگین آن به‌عنوان امتیاز نهایی در نظر گرفته می‌شود که در (۴) مطرح گردیده است. در این رابطه w بیانگر یک واژه از پرس‌وجو یا سند است

$$sim(Query, Doc) = Average(\cos_{w \in Query}(w, Centroid_{Doc})) \qquad (۴)$$

۴-۲ رابطه QrDocCentroidSim

در (۵)، علاوه بر محاسبه نقطه مرکزی سند در فاز برون‌خط، نقطه مرکزی واژه‌های پرس‌وجو در فاز برخط نیز محاسبه می‌گردد. سپس شباهت کسینوسی بین بردار نقطه مرکزی سند و پرس‌وجو محاسبه و به‌عنوان امتیاز آن سند در نظر گرفته می‌شود

$$Sim(Query, Doc) = (Centroid_{Query}, Centroid_{Doc}) \qquad (۵)$$

۴-۳ رابطه ImprovedMaxSim

در رابطه بیشترین شباهت [۲۶] به‌ازای هر واژه‌ی پرس‌وجو، بردار معنایی مربوط از مدل درون‌سازی BERT استخراج می‌گردد. سپس شباهت کسینوسی آن واژه با هر واژه از سند که بردار آن واژه نیز از روی مدل درون‌سازی، استخراج شده محاسبه می‌گردد. در مرحله بعد، بیشترین

1. Centroid

شباهت آن واژه پرس‌وجو با واژه‌های سند بعد از محاسبه، ذخیره و در idf واژه‌ی پرس‌وجو ضرب می‌شود. مجموع این حاصل‌ضرب به‌ازای تک‌تک واژه‌های پرس‌وجو، محاسبه گردیده و نهایتاً نسبت آن بر مجموع idf واژه‌های پرس‌وجو به‌دست می‌آید و به‌عنوان امتیاز اول در نظر گرفته می‌شود. سپس همین محاسبات به‌ازای هر واژه سند و تمامی واژه‌های پرس‌وجو محاسبه می‌گردد و بیشترین شباهت کسینوسی هر واژه سند و تمام واژه‌های پرس‌وجو، محاسبه و در idf آن واژه ضرب می‌گردد. مجموع این حاصل‌ضرب به‌ازای تک‌تک واژه‌های سند با واژه‌های پرس‌وجو محاسبه شده و سپس نسبت آن بر مجموع idf واژه‌های سند به‌دست می‌آید و به‌عنوان امتیاز دوم در نظر گرفته می‌شود که در (۶) قابل مشاهده است.

دلیل استفاده از idf، استفاده صحیح از واژگان و حروف اضافه پرتکرار است. بنابراین واژه‌ای که در تعداد اسناد کمتری ظاهر شود نسبت به واژه‌ای که در بیشتر اسناد وجود دارد مانند حروف اضافه، دارای اطلاعات بیشتری است. مقدار idf بر اساس لگاریتم تعداد اسناد نسبت به اسنادی که شامل واژه t هستند محاسبه می‌گردد

$$Sim(Query, Doc) = \partial\left(\frac{\sum_{w\in Query} \max Sim(w, Doc).idf(w)}{\sum_{w\in Query} idf(w)} + \frac{\sum_{w\in Doc} \max Sim(w, Query).idf(w)}{\sum_{w\in Doc} idf(w)}\right), \partial = ۰/۵ \quad (۶)$$

رابطه پیشنهادی ImprovedMaxSim، روش بهبودیافته بیشترین شباهت هر واژه در پرس‌وجو و سند است؛ با این تفاوت که به‌جای حاصل‌ضرب بیشترین شباهت هر واژه با idf آن واژه به‌صورت نوآورانه، نسبت بیشترین شباهت واژه بر مجذور فرکانس واژه در کل اسناد محاسبه شده و همین روند به‌ازای واژه‌های سند و پرس‌وجو ادامه می‌یابد که در (۷) بیان می‌گردد

$$Sim(Query, Doc) = \partial\left(\frac{\sum_{w\in Query} \max Sim(w, Doc).\frac{۱}{df_w^۲}}{\sum_{w\in Query} \frac{۱}{df_w^۲}} + \frac{\sum_{w\in Doc} \max Sim(w, Query).\frac{۱}{df_w^۲}}{\sum_{w\in Doc} \frac{۱}{df_w^۲}}\right), \partial = ۰/۵ \quad (۷)$$

در رابطه ImprovedMaxSim به‌جای محاسبه idf از نسبت یک بر مربع فرکانس اسناد استفاده گردیده است. دلیل این کار، درون‌سازی همه واژگان در فضای برداری است و این امر در خصوص حروف اضافه هم صدق می‌کند. در درون‌سازی واژگان موجود در هر جمله، حروف اضافه هم مشاهده می‌شود؛ پس آموزش مدل بر مبنای واژگان و حروف اضافه هم صورت می‌گیرد. با مشاهده واژگان آموزش‌یافته از طریق مدل‌ها می‌توان دریافت که واژگان اضافه به‌خوبی تشخیص داده شده و در مکانی نزدیک به یکدیگر قرار گرفته‌اند. از آنجا که این واژگان همه حروف اضافه هستند، در تشخیص بردار معنایی واژگان مشابه خود به‌خوبی عمل می‌کنند. اما در صورتی که هدف از این کار بررسی شباهت میان دو عبارت باشد می‌تواند منجر به تأثیر منفی ناشی از وزن‌های مرتبط با حروف اضافه گردد. به عبارت دیگر در یک دنباله کوتاه از پرس‌وجو و سند، تشخیص واژگان کلیدی و افزایش وزن آنها می‌تواند در کیفیت رتبه‌بندی تأثیرگذار باشد. در واقع به‌دلیل احتمال تأثیر وزن واژگان غیرضروری بر سایر واژگان کلیدی، با کاهش شدیدتر وزن آنها سعی در افزایش دقت و بهبود روند رتبه‌بندی گردیده است.

٤-٤ رابطه UnCommImpMaxSim

رابطه (۸) به محاسبه‌کردن بیشترین شباهت بین واژه‌های پرس‌وجو و سند و واژه‌های غیرمشترک می‌پردازد. در واقع در این روش علاوه بر محاسبه ImprovedMaxSim واژه‌های پرس‌وجو و سند، مشابه (۷) از ImprovedMaxSim واژه‌های غیرمشترک بین پرس‌وجو و سند نیز استفاده می‌شود. در این روش یک بار الگوریتم بهبودیافته بیشترین شباهت میان پرس‌وجو و سند محاسبه می‌گردد و به‌عنوان امتیاز اول در نظر گرفته می‌شود. سپس واژه‌های مشترک در سند و پرس‌وجو حذف می‌شوند و بر روی واژه‌های باقیمانده، مجدداً الگوریتم بهبودیافته بیشترین شباهت محاسبه می‌گردد و به امتیاز اول اضافه می‌شود. به عبارت دیگر، حذف واژه‌های مشترک و محاسبه‌ی شباهت برداری واژگان غیرمشترک، می‌تواند منجر به کشف الگوهای جدید در اسناد گردد.

$$\begin{aligned} Sim(Query, Doc) &= \\ &\max Sim(Query, Doc) + \partial(Query', Doc') = \\ &\partial\left(\frac{\sum_{w\in Query} \max Sim(w, Doc).\frac{۱}{df_w^۲}}{\sum_{w\in Query} \frac{۱}{df_w^۲}} + \frac{\sum_{w\in Doc} \max Sim(w, Query).\frac{۱}{df_w^۲}}{\sum_{w\in Doc} \frac{۱}{df_w^۲}}\right) + \\ &\partial\left(\frac{\sum_{w\in Query'} \max Sim(w, Doc').\frac{۱}{df_w^۲}}{\sum_{w\in Query'} \frac{۱}{df_w^۲}} + \frac{\sum_{w\in Doc'} \max Sim(w, Query').\frac{۱}{df_w^۲}}{\sum_{w\in Doc'} \frac{۱}{df_w^۲}}\right), \partial = ۰/۵ \end{aligned} \quad (۸)$$

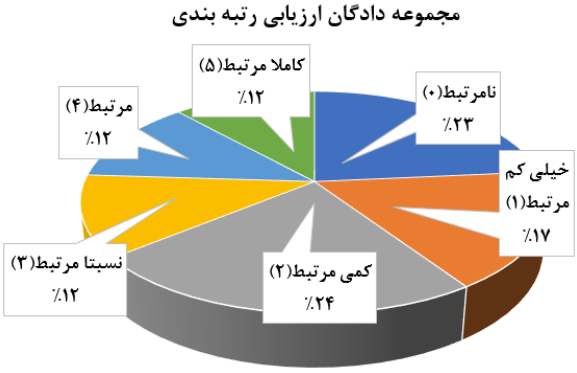
در این رابطه از عبارات پرس‌وجو و سند در بخش اول و از عبارت پرس‌وجو و سند با واژگان غیرمشترک برای بخش دوم استفاده می‌گردد.

شبه‌کد (۷) در شکل ۲ به‌عنوان رابطه پیشنهادی این مقاله بیان گردیده است.

تابع ImprovedMaxSim، دو ورودی Query و Doc را که به‌ترتیب نشان‌دهنده پرسش و سند هستند، دریافت می‌کند. این تابع ابتدا با استفاده از درون‌سازی BERT، حداکثر شباهت کسینوسی هر واژه را در عبارت پرس‌وجو با تمام واژگان موجود در سند محاسبه می‌نماید. سپس با ضرب حداکثر شباهت کسینوسی با معکوس مربع فراوانی سند، شباهت بین عبارت پرس‌وجو و سند و همچنین فراوانی معکوس سند و پرس‌وجو را محاسبه می‌کند. نهایتاً با جمع‌شدن امتیاز شباهت عبارت پرس‌وجو و سند، امتیاز شباهت نهایی محاسبه می‌شود.

```
function ImprovedMaxSim(Query, Doc):
    N = total number of documents
    sim_query_doc = 0
    sim_doc_query = 0
    for each word_w in Query:
        max_sim_w_doc = maximum cosine similarity of
        | word_w with all words in Doc using BERT embedding.
        df_w = number of documents containing word_w
        sim_query_doc += max_sim_w_doc * (1 / (df_w ** 2))
    for each word_w in Doc:
        max_sim_w_query = maximum cosine similarity of
        | word_w with all words in Query using BERT embedding.
        df_w = number of documents containing word_w
        sim_doc_query += max_sim_w_query * (1 / (df_w ** 2))
    idf_query = log(N / number of documents containing any word in Query)
    idf_doc = log(N / number of documents containing any word in Doc)
    sim_query_doc *= (1 / (sum of (1 / (df_w ** 2))
    |                 |   for each word_w in Query)) + idf_query
    sim_doc_query *= (1 / (sum of (1 / (df_w ** 2))
    |                 |   for each word_w in Doc))+ idf_doc
    similarity = (0.5 * (sim_query_doc + sim_doc_query))
    return similarity
```

شکل ۲: شبه‌کد رابطه ImptovedMaxSim.



شکل ۳: مجموعه دادگان مورد استفاده جهت ارزیابی فرمول‌های رتبه‌بندی.

۵- مجموعه دادگان

مجموعه دادگان جهت رتبه‌بندی اسناد و ارزیابی مدل‌ها در این مرحله، شامل جفت پرس‌وجو و مجموعه اسناد واقعی است. این مجموعه دادگان، شش‌برچسبه است و به‌ازای هر پرس‌وجو، متوسط ۱۰ عنوان سند در نظر گرفته می‌شود که با عددهای صفر تا پنج بر مبنای میزان ارتباط سند به پرس‌وجو توسط تیم خبره، برچسب‌گذاری می‌گردد. این مجموعه دادگان شامل ۵۰۰ پرس‌وجو است و جزئیاتش در شکل ۳ قابل مشاهده است. این مجموعه شامل ۲۰۰ پرس‌وجوی مورد استفاده در [۲۳] می‌باشد که با افزودن ۳۰۰ پرس‌وجوی جدید، ارتقا یافته است. این اسناد بر اساس تکنیک TF.IDF به دقت ۰٫۶۴ دست می‌یابند. در ادامه، رتبه‌بندی اسناد بر مبنای مدل‌های درون‌یابی و استفاده از فرمول‌های رتبه‌بندی مورد ارزیابی قرار می‌گیرد.

برای ارزیابی و مقایسه کیفیت رتبه‌بندی در بازیابی اطلاعات، از معیار nDCG مطابق (۹) استفاده می‌گردد. در این رابطه که برای n نتیجه اول محاسبه می‌شود، r_j بیانگر درجه ارتباط سند j با پرس‌وجوی مربوط است [۲۷]

$$nDCG@n = \sum_{j=1}^{n} \frac{2^{r_j} - 1}{\log(1 + j)} \qquad (۹)$$

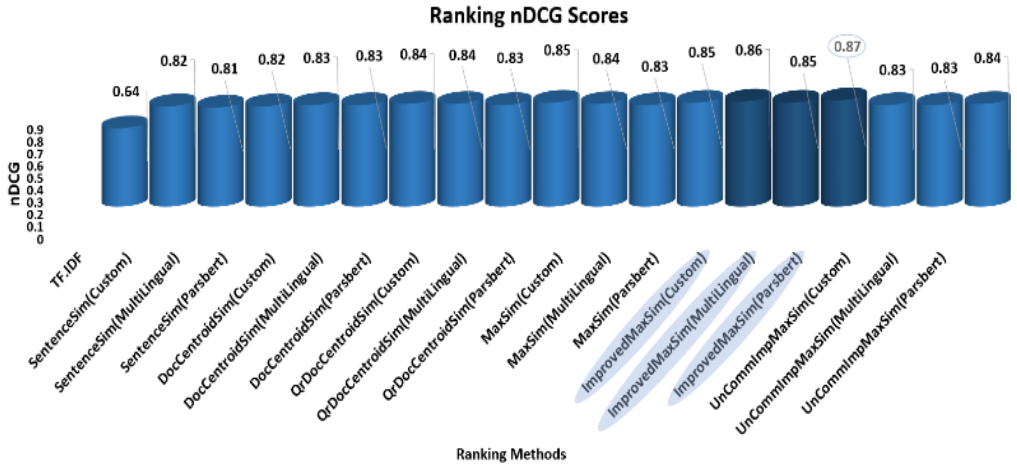
nDCG دو عامل اصلی را در نظر می‌گیرد: ارتباط سند با پرس‌وجوی کاربر و موقعیت آن سند در رتبه‌بندی. برای محاسبه معیار nDCG، میزان ارتباط هر سند به پرس‌وجوی مربوط برچسب‌گذاری می‌گردد که نشان از میزان تطبیق سند با پرس‌وجوی کاربر است. سپس DCG با جمع امتیازات برترین سندها محاسبه می‌شود و با درنظرگرفتن موقعیت سند به‌دست می‌آید. nDCG با تقسیم DCG به‌دست‌آمده از رتبه‌بندی بر DCG ایده‌آل به دست می‌آید. امتیاز بالاتر nDCG به معنای رتبه‌بندی دقیق‌تر است.

۶- نتایج رتبه‌بندی با استفاده از درون‌سازی BERT

نتایج حاصل از ارزیابی روابط رتبه‌بندی به‌ازای سه مدل درون‌سازی BERT در شکل ۴ قابل بررسی است. رتبه‌بندی بر مبنای بردار درون‌سازی جملات ورودی و بدون درنظرگرفتن بردار هر واژه به‌صورت مجزا تحت عنوان رابطه SentenceSim در بهترین حالت برابر با ۰٫۸۲ به‌دست می‌آید. این دقت مبتنی بر بردار معنایی تولیدشده از مدل BERT به‌ازای جملات ورودی است؛ اما در صورت بهره‌گیری از روابط پیشنهادی، دقت به‌خوبی بهبود می‌یابد که در ادامه به آن می‌پردازیم.

بهترین دقت کسب‌شده از رابطه DocCentroidSim، مربوط به مدل پارس برت آموزش‌یافته بر مبنای معماری شکل ۱ است که برابر با ۰٫۸۴ به‌دست می‌آید. در راستای همین ارزیابی، مدل برت سفارشی و برت چندزبانه به دقت ۰٫۸۳ دست می‌یابند که قابل توجه است. روابط QrDocCentroidSim و MaxSim با بهبود دقت رتبه‌بندی تا ۱٪ در بهترین حالت به دقت ۰٫۸۵ دست پیدا خواهند کرد. در رابطه QrDocCentroidSim می‌توان به تفاوت عملکرد نقطه مرکزی واژگان و خود واژه پرداخت؛ زیرا از روی بردار یک واژه می‌توان آن واژه را در فضای n بُعدی مشاهده کرد و نزدیک‌ترین واژه‌ها به آن واژه را برگزید؛ اما از روی بردار نقطه مرکزی به‌دست‌آمده از واژه‌های یک پرس‌وجو، به دلیل آنکه هیچ واژه‌ای در n امین بُعد با این بردار همخوانی ندارد می‌توان دریافت که این بردار به‌جای اشاره به یک واژه مشخص، به یک موضوع اشاره دارد؛ لذا در صورت استخراج واژه‌های نزدیک به این بردار به واژگانی دست می‌یابیم که از نظر موضوعی به هم نزدیک هستند. در نتیجه، محاسبه شباهت میان نقطه مرکزی واژه‌های پرس‌وجو و سند به نتایج دقیق‌تری می‌رسد که قابل توجه است.

در رابطه MaxSim جزئیات بیشتری در خصوص هر واژه و میزان شباهت آن با سایر واژه‌ها در نظر گرفته می‌شود. در این روش، یک بار به ازای هر واژه پرس‌وجو و واژه‌های سند، بیشترین شباهت به‌دست می‌آید و به‌صورت مشابه همین روال برای هر واژه سند و واژه‌های پرس‌وجو تکرار



شکل۴: نتایج رتبه‌بندی بر مبنای روابط پیشنهادی.

می‌شود. در این رابطه، idf واژه‌ها نیز در نظر گرفته می‌شود تا واژه‌های پرتکرار کم‌اهمیت، تأثیر کمتری در امتیاز حاصل از این روش داشته باشند. دلیل افزایش دقت در این رابطه، ارزیابی میزان شباهت هر واژه از عبارت اول با هر یک از واژگان عبارت دوم و استفاده از بیشترین شباهت در رتبه‌بندی است. با تکرار این فرایند برای عبارت دوم در ازای عبارت اول، شباهت تمام واژگان به‌درستی محاسبه می‌گردد و بیشترین شباهت ناشی از هر واژه به‌صورت مجزا در نظر گرفته می‌شود که تأثیر خوبی در افزایش دقت در بر داشته است.

رابطه پیشنهادی ImprovedMaxSim، روش بهبودیافته رابطه MaxSim است که با ایده‌ای جدید و بهره‌گیری از نسبت یک بر مربع فرکانس اسناد به‌جای idf می‌تواند در بهترین حالت به افزایش دقت تا ۰/۸۷ بر مبنای مدل پارس برت نائل گردد. این راهکار با هدف کاهش حداکثری وزن واژه‌های اضافی و پرتکرار و استخراج واژگان کلیدی در درک بهتر مفهوم هر عبارت مورد استفاده قرار می‌گیرد. واضح است که تأثیر واژه‌ای که در تعداد اسناد کمتری ظاهر شود نسبت به واژه‌ای که در بیشتر اسناد وجود دارد بیشتر است. به عنوان نمونه می‌توان به واژگان اضافه اشاره کرد. برخی از روش‌های رتبه‌بندی با درنظرگرفتن تکرار آن واژگان در اسناد مختلف و از طریق idf به شناسایی و در نتیجه کاهش تأثیر آنها در امتیازات خود می‌پردازند. از آنجا که در این مقاله از رتبه‌بندی اسناد به‌صورت مستقیم بر مبنای بردار معنایی واژگان استفاده گردیده است باید تأثیر واژگان اضافه و بردارهای معنایی آنها بر رتبه‌بندی در نظر گرفته شود. به همین دلیل در این رابطه از نسبت یک بر مربع فرکانس اسناد استفاده شد. دلیل این کار، درون‌سازی همه واژگان در فضای برداری است و این امر در خصوص حروف اضافه هم صدق می‌کند. با مشاهده واژگان آموزش‌یافته از طریق مدل‌ها می‌توان دریافت که واژگان اضافه به خوبی تشخیص داده شده و در مکانی نزدیک به یکدیگر قرار گرفته‌اند و کاهش تأثیر آنها در امتیازدهی به کسب دقت بالاتر کمک می‌کند.

با توجه به تأثیر حداکثری رابطه ImprovedMaxSim در نتایج و با هدف تأکید بیشتر روی واژه‌های غیرمشترک بین سند و پرس‌وجو، رابطه UnCommImpMaxSim مورد بررسی قرار می‌گیرد. در مرحله اول بیشترین شباهت بین واژه‌های پرس‌وجو و سند محاسبه می‌گردد و سپس در مرحله دوم با نادیده‌گرفتن واژه‌های مشترک بین پرس‌وجو و سند، مجدداً بیشترین شباهت بین واژه‌های غیرمشترک سند و پرس‌وجو محاسبه و مجموع آن به‌عنوان امتیاز نهایی ثبت می‌گردد. به عنوان مثال اگر پرس‌وجو با محتوای «علائم سکته قلبی» و سند با محتوای «همه چیز در مورد سکته قلبی» در نظر گرفته شود، علاوه بر محاسبه بیشترین شباهت بین سند و پرس‌وجوی مربوط، بیشترین شباهت بین «علائم» و «همه‌چیز در مورد» نیز جداگانه محاسبه می‌گردد. اما در صورتی که سند دیگری با محتوای «علائم سکته مغزی» موجود باشد می‌توان با بررسی بیشتر دریافت که شباهت بین پرس‌وجو و سند اول در واژه‌های غیرمشترک، کمتر از سند دوم است. این مسئله به این دلیل رخ می‌دهد که دو واژه «قلبی» و «مغزی»، ارتباط معنایی قوی‌تری نسبت به ارتباط واژه‌های «علائم» و «همه‌چیز در مورد» دارند؛ لذا دقت در این رابطه نسبت به رابطه ImprovedMaxSim کاهش می‌یابد و در بهترین حالت به دقت ۰/۸۴ می‌رسد.

با مقایسه نتایج رتبه‌بندی اسناد می‌توان دریافت که نتایج بر مبنای بردارهای درون‌سازی استخراج‌شده از مدل‌های درون‌سازی و با بهره‌گیری از روابط رتبه‌بندی پیشنهادی از دقت بالایی برخوردار هستند. نکته دیگر در بهبود نتایج مدل برت سفارشی نسبت به مدل چندزبانه است. به‌عنوان مثال دقت مدل برت سفارشی بر پایه رابطه ImprovedMaxSim برابر با ۰/۸۶ به‌دست می‌آید که نسبت به مدل برت چندزبانه با کسب دقت ۰/۸۵ بهبود داشته است.

۷- نتیجه‌گیری

در این مقاله یک رویکرد رتبه‌بندی جدید پیشنهاد می‌گردد که با استخراج بردارهای معنایی از طریق مدل زبانیBERT منجر به افزایش دقت رتبه‌بندی اسناد می‌گردد. رابطه ImprovedMaxSim به‌عنوان فرمول رتبه‌بندی پیشنهادی با بررسی بیشترین شباهت هر واژه پرس‌وجو با کل واژگان سند و درنظرگرفتن تکرار آن واژه در اسناد و بالعکس، به رتبه‌بندی اسناد بر مبنای پرس‌وجوی کاربر می‌پردازد. این رابطه با کسب دقت ۰/۸۷ که بالاترین دقت به‌دست‌آمده نسبت به سایر روش‌های رتبه‌بندی مورد بررسی در این ارزیابی است، به‌عنوان فرمول رتبه‌بندی مناسب در این مقاله معرفی و پیشنهاد می‌گردد.

به طور کلی، ترکیب مدل درون‌سازی BERT و فرمول‌های رتبه‌بندی پیشنهادی، رویکردی موفق برای افزایش دقت رتبه‌بندی در سیستم‌های بازیابی اطلاعات است. این تحقیق به پیشرفت تکنیک‌های رتبه‌بندی اسناد کمک می‌کند و پایه‌ای محکم را برای پیشرفت‌های آینده در این زمینه فراهم می‌نماید. همچنین امکانات جدیدی را برای بهبود تجربیات جستجو و حصول اطمینان از دریافت نتایج جستجوی بسیار مرتبط و دقیق فراهم می‌سازد.

مدل BERT به‌دلیل محاسبات فراوان در لایه‌های مختلف، هزینه بالایی را در بر دارد. همچنین در زمان استنتاج، نیازمند محاسبات فراوان است و مشابه سایر روش‌ها در متون طولانی از دقت کافی برخوردار نیست. در راستای حل این مسئله می‌توان از ترکیب الگوریتم‌های درون‌سازی با پیکربندی سبک‌تر و بهره‌گیری از محاسن هر مدل در بهبود رتبه‌بندی استفاده نمود که به‌عنوان کارهای آینده در نظر گرفته می‌شود. از کارهای دیگری که می‌توان در آینده به آن پرداخت، بهره‌گیری از درون‌سازی واژگان در رتبه‌بندی اسناد مبتنی بر گراف است. بردارهای درون‌سازی واژگان را می‌توان به‌عنوان دانش پس‌زمینه در رویکردهای رتبه‌بندی مبتنی بر گراف در نظر گرفت تا به درک اطلاعات ساختاری و معنایی منجر گردد و در رتبه‌بندی بهتر اسناد مورد استفاده قرار گیرد.

مراجع

[1] Y. Yum, *et al.*, "A word pair dataset for semantic similarity and relatedness in Korean medical vocabulary: reference development and validation," *JMIR Medical Informatics*, vol. 9, no. 6, Article ID: e29667, Jun. 2021.

[2] E. Hindocha, V. Yazhiny, A. Arunkumar, and P. Boobalan, "Short-text semantic similarity using GloVe word embedding," *International Research J. of Engineering and Technology*, vol. 6, no. 4, pp. 553-558, Apr. 2019.

[3] J. Zhang, Y. Liu, J. Mao, W. Ma, and J. Xu, "User behavior simulation for search result re-ranking," *ACM Trans. on Information Systems*, vol. 41, no. 1, Article ID: 5, 35 pp., Jan. 2023.

[4] V. Zosimov and O. Bulgakova, "Usage of inductive algorithms for building a search results ranking model based on visitor rating evaluations," in *Proc. IEEE 13th Int. Scientific and Technical Conf. on Computer Sciences and Information Technologies, CSIT'18,* pp. 466-469, Lviv, Ukraine, 11-14 Sept. 2018.

[5] B. Mitra and N. Craswell, *Neural Models for Information Retrieval*, arXiv preprint arXiv:1705.01509, vol. 1, 2017.

[6] V. Gupta, A. Dixit, and S. Sethi, "A comparative analysis of sentence embedding techniques for document ranking," *J. of Web Engineering*, vol. 21, no. 7, pp. 2149-2186, 2022.

[7] J. Pennington, R. Socher, C. Ma, and C. Manning, "GloVe: global vectors for word representation," in *Proc. Conf. on Empirical Methods in Natural Language Processing, EMNLP'14,* pp. 1532-1543, Doha, Qatar, 25-29 Oct. 2014.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dea, "Efficient estimation of word representations in vector space," in *Proc. In. Conf. on Learning Representations, ICLR'13,* 12 pp., Scottsdale, AZ, USA, 2-4 May 2013.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017.

[10] M. E. Peters, et al., "Deep contextualized word representations," in Proc. Conf. of the North American Chapter of the Association of Computational Linguistics, NAACL-HLT'18, 11 pp., New Orleans, LA, USA, 1-6 Jun. 2018.

[11] J. Devlin, M. W. Chang, and K. L. Kristina, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. of the North American Chapter of the Association of Computational Linguistics, NAACL-HLT'18,* 16 pp., New Orleans, LA, USA, 1-6 Jun. 2018.

[12] T. Brown, *et al.*, "Language models are few-shot learners," in *Proc. 34th Conf. on Neural Information Processing Systems, NeurIPS'20,* 25 pp., Vancouver, Canada, 6-12 Dec. 2020.

[13] P. Sherki, S. Navali, and R. Inturi, "Retaining semantic data in binarized word embedding," in *Proc. IEEE 15th Int. Conf. on Semantic Computing, ICSC'21,* pp. 130-133, Laguna Hills, CA, USA, 27-29 Jan. 2021.

[14] L. Shaohua, C. Tat-Seng, Z. Jun, and C. Miao, *Generative Topic Embedding: A Continuous Representation of Documents (Extended Version with Proofs)*, arXiv preprint arXiv:1606.02979, vol. 1, 2016.

[15] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana, "A dual embedding space model for document ranking," in *Proc. 25th Int. Conf. Companion on World Wide Web, WWW'16,* 10 pp., Montreal, Canada, 11-15 Apr. 2016.

[16] M. Dehghani, H. Zamani, A. Severyn, and J. Kamps, "Neural ranking models with weak supervision," in *Proc. of the 40th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR '17,* pp. 65-74, Tokyo, Japan, 7-11 Aug. 2017.

[17] C. Xiong, Z. Dai, and J. Callan, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proc. of the 40th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval,* pp. 55-64, Tokyo, Japan, 7-11 Aug. 2017.

[18] R. Brochier, A. Guille, and J. Velcin, "Global vectors for node representations," in Proc. *ACM World Wide Web Conf., WWW'19, San Francisco*, pp. 2587-2593, San Francisco, CA, USA, 13-17 May 2019.

[19] A. Gourru and J. Velcin, "Gaussian embedding of linked documents from a pretrained semantic space," in *Proc. 29th Int. Joint Conf. on Artificial Intelligence, IJCAI'20,* pp. 3912-3918, Yokohama, Japan, 8-10 Jan. 2021.

[20] R. Menon, J. Kaartik, and K. Nambiar, "Improving ranking in document based search systems," in *Proc. 4th Int. Conf. on Trends in Electronics and Informatics, ICOEI'20,* pp. 914-921, Tirunelveli, India, 15-17 Jun. 2020.

[21] J. Li, C. Guo, and Z. Wei, "Improving document ranking with relevance-based entity embeddings," in *Proc. 8th Int. Conf. on Big Data and Information Analytics, BigDIA'22, China*, pp. 186-192, Guiyang, China, 24-25 Aug. 2022.

[22] S. Han, X. Wang, M. Bendersky, and M. Najork, *Learning-to-Rank with BERT in TF-Ranking*, Google Research Tech Report, 2020.

[۲۳] ش. بستان، ع. زارع بیدکی و م. ر. پژوهان، "درون‌سازی معنایی واژه‌ها با استفاده از BERT روی وب فارسی،" *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب- مهندسی کامپیوتر*، سال ۲۱، شماره ۲، صص. ۱۰۰-۸۹، تابستان ۱۴۰۲.

[24] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: transformer-based model for Persian language understanding," *Neural Processing Letters*, vol. 53, pp. 3831-3847, 2021.

[25] D. Yang and Y. Yin, "Evaluation of taxonomic and neural embedding methods for calculating semantic similarity," *Natural Language Engineering*, vol. 28, no. 6, pp. 733-761, Nov. 2022.

[26] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. 21st National Conf. on Artificial Intelligence,* vol. 1, pp. 775-780, Boston, MA, USA, 16-20 Jul. 2006.

[27] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. on Information Systems*, vol. 20, no. 4, pp. 422-446, Oct. 2002.

شکوفه بستان دکتری مهندسی کامپیوتر با گرایش نرم‌افزار از دانشگاه یزد است. او در حال حاضر به عنوان مدرس در دانشکده مهندسی کامپیوتر دانشگاه یزد و همچنین به عنوان توسعه‌دهنده نرم‌افزار در یک شرکت برجسته جستجوی ابری فعالیت دارد. زمینه‌های تحقیقاتی مورد علاقه ایشان شامل یادگیری عمیق، بازیابی معنایی اطلاعات و تحلیل معنایی شبکه‌های اجتماعی است.

علی‌محمد زارع‌بیدکی تحصیلات خود را در مقطع کارشناسی در سال ۱۳۷۸ از دانشگاه صنعتی اصفهان و مقاطع کارشناسی ارشد و دکتری کامپیوتر را به‌ترتیب در سال‌های ۱۳۸۱ و ۱۳۸۸ از دانشکده فنی دانشگاه تهران به پایان رسانده است و هم‌اکنون عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان شامل بازیابی اطلاعات، موتورهای جستجو، رتبه‌بندی و پردازش زبان‌های طبیعی است.

محمدرضا پژوهان استادیار گروه مهندسی کامپیوتر دانشگاه یزد است. او دکترای خود را در بخش علوم کامپیوتر از دانشگاه ساینس مالزی (USM) و دانشگاه ملی سنگاپور (NUS) اخذ کرده است. ایشان فارغ‌التحصیل کارشناسی و کارشناسی ارشد مهندسی کامپیوتر از دانشگاه صنعتی شریف است. علایق تحقیقاتی ایشان شامل پایگاه داده، داده‌کاوی، علوم داده و حفظ حریم خصوصی در انتشار داده‌هاست.