IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2023, accepted 3 July 2023, date of publication 14 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3295776

SURVEY

Information Retrieval: Recent Advances and Beyond

KAILASH A. HAMBARDE AND HUGO PROENÇA, (Senior Member, IEEE)
Department of Computer Science, Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal

Corresponding author: Kailash A. Hambarde (kailas.srt@gmail.com)

This work was supported in part by Adaptative Designed Clinical Pathways (AddPath) Project Fundacao para a Ciencia e a Tecnologia (FCT)/Ministry of Science, Technology and Higher Education (MCTES) through National Funds under Grant CENTRO-01-0247-FEDER-072640 and Grant LISBOA-01-0247-FEDER-072640, and in part by EU758 Funds under Project UIDB/50008/2020.

ABSTRACT This paper provides an extensive and thorough overview of the models and techniques utilized in the first and second stages of the typical information retrieval processing chain. Our discussion encompasses the current state-of-the-art models, covering a wide range of methods and approaches in the field of information retrieval. We delve into the historical development of these models, analyze the key advancements and breakthroughs, and address the challenges and limitations faced by researchers and practitioners in the domain. By offering a comprehensive understanding of the field, this survey is a valuable resource for researchers, practitioners, and newcomers to the information retrieval domain, fostering knowledge growth, innovation, and the development of novel ideas and techniques.

INDEX TERMS First-stage retrieval, information retrieval, second-stage retrieval.

I. INTRODUCTION

Currently, Information Retrieval (IR) holds significant importance in people's daily lives due to its integration in various useful functions alike as internet browsing, question-answering systems, personal assistants, chatbots, and digital libraries. The primary objective is to recognize and retrieve information that is associated with the user's request. Since multiple records may be relevant, the results are frequently ranked according to their similarity score to the user's query. At the beginning of the IR field, traditional text retrieval systems predominantly rely on matching terms between documents and queries. However, these term-based retrieval systems have drawbacks, such as polysemy, synonymy, and lexical gaps, which can limit their effectiveness [1]. Recently, the field of Natural Language Processing (NLP) has undergone significant advancements due to the increased availability of large labeled datasets and enhanced computing power, which has allowed researchers to employ deep learning methods for various purposes. These techniques have been utilized to enhance traditional text retrieval systems and address the limitations of term-based retrieval techniques. However, applying these techniques requires substantial amounts of data and computational resources. As a result, researchers are constantly developing more advanced deep learning algorithms to meet these demands and achieve better results in NLP tasks [2]. With the use of these advanced deep learning algorithms, the performance of IR systems has been significantly improved, leading to more precise and efficient retrieval of information for end-users. Some of the advancements in deep learning techniques that have been employed in IR include neural network architectures such as convolutional neural networks [3] and recurrent neural networks [4], in addition to transfer learning and pre-training techniques [5]. These methods have improved text data representation and enhanced the IR system's understanding of natural language queries. Moreover, attention-based mechanisms like the Transformer architecture [6] have been implemented to enhance the capability of IR systems to focus on critical parts of the query and documents for matching purposes. Further, adopting pre-trained language models, such as BERT [5] and GPT-2 [7], has demonstrated the ability to improve IR systems performance by providing superior cognition of the semantics and context of natural

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Abid.

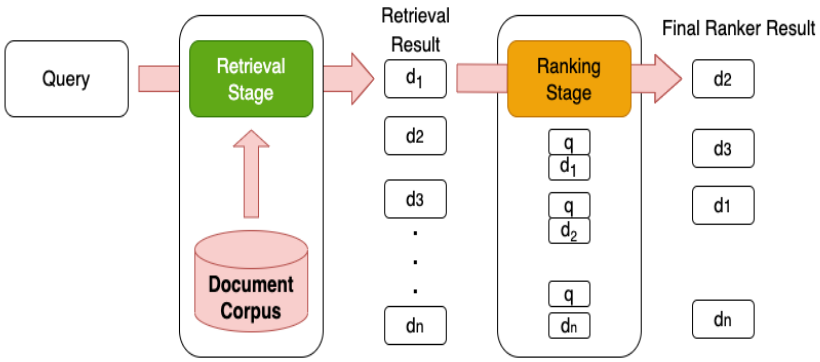

FIGURE 1. Overview of modern Information Retrieval system.

TABLE 1. Categorization of Previous Surveys on Information Retrieval Models into Retrieval and Ranking Categories.

References	Year	Retrieval	Ranking
Hang Li et al. [17]	2014	✗	✓
Onal et al. [14]	2018	✓	✗
Jiafeng et al. [13]	2020	✗	✓
Yates Andrew et al. [2]	2020	✗	✓
Jimmy Lin et al. [18]	2020	✓	✓
Y Cai et al. [16]	2021	✓	✗
This survey	2023	✓	✓

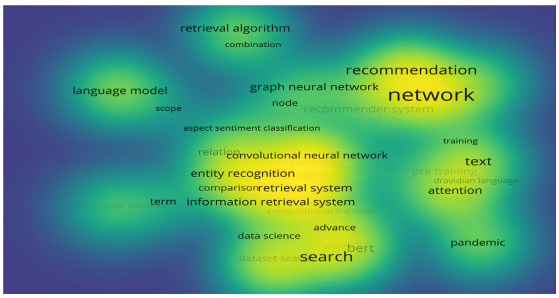

FIGURE 2. Term map of the information retrieval. Colors indicate the recent term density extracted from survey papers.

language queries and documents. Similarly, researchers also concentrated on integrating external knowledge to enhance the relevance of the retrieved information. One approach is to incorporate knowledge graph embeddings [8] into the IR process, which can aid in linking the query and documents to relevant entities and concepts, thereby producing more accurate results. Moreover, the use of multi-modal information retrieval, which involves combining text, image, and audio information, has demonstrated the ability to enhance the performance of IR systems [9].

In general, developing deep learning techniques has significantly improved the performance of information retrieval (IR) systems, enabling them to handle the complexity of natural language queries. This is due to the availability of large labeled datasets and high computational power, yet there is still room for development and investigation in this area.

This survey covers both the first and second stages; models covered in our discussion include those built using words, semantic retrieval, and neural methods. The density term plot depicted in Figure (2) offers a graphical representation of the relative frequency of keywords in the surveyed literature, providing valuable insights into current research trends in the field. The table (1) categorizes previous surveys into retrieval and ranking categories. The table provides a useful summary of the research trends in the field of information retrieval and the topics covered in previous surveys.

To the best of our knowledge, this is the second survey that covers both first-stage retrieval and second-stage ranking models. This survey focuses on papers published in major conferences and journals in the fields of deep learning, natural language processing, and information retrieval from 2013 to June 2022. Table (1) summarizes previous surveys on neural models for Information Retrieval. We also include a comprehensive mind-map, Fig. (3), visually representing the techniques and methods discussed in the subsequent sections. The mind map is organized into two main stages—Retrieval and Ranker—covering conventional retrieval, dense retrieval methods, sparse retrieval methods, hybrid retrieval methods, learning to rank, and deep learning-based ranking models.

The organization of this paper is as follows. Section II introduces the two typical stages of information retrieval models and provides background knowledge, including problem formalization. Section III discusses first-stage retrieval. Section IV discusses second-stage ranking. Section V introduces the SOTA benchmark datasets, while Section VI covers current challenges and future directions. Finally, we conclude the survey in Section VII.

II. INFORMATION RETRIEVAL: OVERVIEW

This part initially discusses the two-stage process, retrieval, and ranking, followed by formulating dense retrieval.

A. DENSE TEXT: RETRIEVAL AND RANKING

Modern information retrieval aims to provide users with the most relevant information to their queries. This objective is accomplished in two stages: retrieval and ranking, as depicted in Figure (1). During the retrieval stage, a collection of initial documents which are potentially relevant to the query is retrieved. After the relevance of these documents is

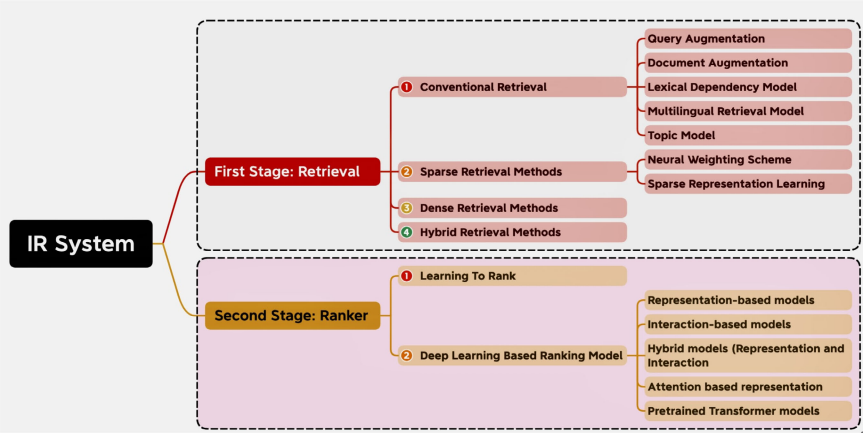

FIGURE 3. Overview of Information Retrieval and Ranking Techniques: A Mindmap.

reassessed based on their similarity scores, the ranking is adjusted accordingly. This is done by using various algorithms and models like vector space model [10], Boolean model, Latent Semantic Indexing [11], Latent Dirichlet Allocation [12], and recent techniques such as pre-trained models like BERT [5].

Within the second stage, ranking, the main goal is to adjust the ranking of the initially retrieved documents based on the relevance score. The ranking process typically employs different models than those used in the retrieval stage, as the primary focus is improving the results' effectiveness rather than their efficiency. Traditional models such as BM25 [18] are used as initial retrievers as they prioritize efficiently recalling relevant documents from a massive document pool. Traditional ranking models encompass a range of techniques; such models include RankNet [19] and [20] for learning to rank and DRMM [21] and Duet [22] for neural models. These models leverage various techniques, including reinforcement learning [23], contextual embeddings [24], and attention mechanisms [25], to learn how to rank documents based on the user's query criteria.

B. FORMULATION

Formally, let q represent a query and d_i symbolize a text from an extensive text collection $D = d_{i=1}^{m}$, comprised of m documents. Text retrieval aims to provide a ranked list of n highly relevant texts $L = [d_1, d_2, \ldots, d_n]$ based on the relevance scores generated by a retrieval model. Technically, we can employ either sparse retrieval models or dense retrieval models to create the retriever. Dense retrieval is characterized by representing queries and texts as dense vectors, enabling the computation of relevance scores using a similarity function between these vectors. This can be expressed as:

$$\text{Rel}(q, d) = f_{\text{sim}}(\phi(q), \psi(d)), \tag{1}$$

where $\phi(\cdot) \in \mathbb{R}^l$ and $\psi(\cdot) \in \mathbb{R}^l$ denote functions that map queries and texts to l-dimensional vectors, respectively. In the context of dense retrieval, $\phi(\cdot)$ and $\psi(\cdot)$ are constructed using neural network encoders, and similarity measurement functions can be used to implement $f_{\text{sim}}(\cdot)$.

III. FIRST STAGE: RETRIEVAL

This section presents a comprehensive literature review of the first-stage retrieval, divided into four categories: conventional, sparse, dense, and hybrid retrieval techniques.

A. CONVENTIONAL RETRIEVAL

Over the past several decades, substantial progress has been made in developing and improving term-based retrieval methods. A diverse array of approaches have been introduced to optimize query and document representations for information retrieval. These techniques often incorporate external resources or capitalize on the inherent information within the collection to enrich the representations. The following section provides a concise overview of some prominent methods in this domain.

1) QUERY AUGMENTATION

The need for query augmentation arises due to the inherent challenges in information retrieval, such as vocabulary mismatch between queries and documents, users' limited ability to express their information needs, and the inherent ambiguity of natural language. To address these challenges, early query augmentation efforts primarily focused on employing query expansion techniques to boost retrieval performance. One approach involved using global models to pinpoint and incorporate concepts relevant to a query [26]. Another strategy harnessed lexical-semantic relationships to identify terms semantically related to the query [27]. These methods exemplify how external resources or the collection itself can be effectively utilized to enrich query representations.

Another effective strategy to enhance query representation is integrating external user feedback into the retrieval process, an extensively researched method. The Rocchio relevance feedback algorithm [29], a well-established local model, is the basis for numerous contemporary relevance feedback techniques. By employing a divergence minimization model, Zhai and Lafferty [30] introduced a technique

for integrating user feedback into information retrieval systems. Cao et al. [31] presented a method for implementing pseudo-relevance feedback (PRF) in these systems. In contrast, Lv and Zhai [32] conducted an analysis of various PRF approaches for information retrieval systems. Furthermore, Zamani et al. [33] proposed an innovative method for incorporating PRF into information retrieval systems using matrix factorization. These methods significantly enhance term-based retrieval approaches, leading to more accurate and relevant search results. Despite query augmentation's benefits, it has some drawbacks, such as the potential for query drift and overfitting. Researchers have explored document augmentation techniques to address these challenges as an alternative approach. In the next section, we will investigate these document augmentation methods and how they have been employed to enhance information retrieval performance.

2) DOCUMENT AUGMENTATION

Document augmentation is an alternate approach to query augmentation conducted on all documents in the corpus. This method supplements every posting list within an inverted index, which has been found to be particularly useful. However, while document augmentation enhances the data richness, it is crucial to avoid over-augmentation that could introduce redundancy, leading to inefficiencies or potential overfitting in the subsequent stages of data processing or model training.

The concept of document expansion, which originated in the field of speech retrieval [133], has since been explored and developed by numerous researchers, highlighting its significance in information retrieval. Multiple studies, such as [38] and [39], have demonstrated the effectiveness of various document augmentation methods. For instance, Kurland and Lee [34] and Liu and Croft [35] investigated the relations between language models, corpus structure, and ad-hoc retrieval. Liu and Croft [35] put forth a document clustering approach, whereas Billerbeck and Zobel [36] conducted a comparative analysis of document augmentation and query augmentation techniques in ad-hoc retrieval. In addition, several researchers have focused on expanding document representations by incorporating related terms to boost retrieval performance. Tao et al. [37] proposed innovative techniques for this purpose, while Agirre et al. [38] explored WordNet, a comprehensive lexical database of English, for document expansion. Alternatively, Efron et al. [39] concentrated on enhancing short text retrieval by adding semantically related terms to each document's representation. Furthermore, Sherman et al. [40] expand document representations using external collections like WordNet. These studies demonstrate the field's evolution, with more recent papers building upon their predecessors' ideas and exploring novel document expansion methods in information retrieval systems.

3) LEXICAL DEPENDENCY MODEL

Traditional methods in information retrieval often treat terms in a document as independent entities without considering their order or relationships. This approach may result in an inaccurate representation of concepts that consist of multiple contiguous words and a less effective capture of relevance that arises from the specific order of terms when matching queries and documents. To overcome these limitations, lexical dependency models have been developed to incorporate term dependencies into their representation functions. By considering the relationships and order of terms, these models can more accurately capture the underlying structure and meaning of the text. This, in turn, enables a more precise representation of concepts and improved matching between queries and documents, ultimately enhancing information retrieval performance. Researchers have studied this; Fagan [41] have significantly contributed to the area of lexical dependency models by attempting to incorporate phrases within the vector space model, treating them as additional dimensions within the representation space. Their study compares the effectiveness of different automatic phrase indexing methods. Another pivotal contribution is by Salton and Buckley [134], which introduced a novel term dependency weighting scheme that considers the relationship between terms in a document. Subsequent research has expanded on these ideas, such as Mitra et al. [42], who propose a new phrase-based retrieval method using a vector space model and term dependency weighting. Song and Croft [43] presented a strategy based on a general language model incorporating term dependency weighting. Probabilistic models have been explored in this context as well. Jones et al. [44] suggested a new probabilistic model that builds on the vector space model and incorporates term dependency weighting. Nallapati and Allan [45] recommend using sentence trees to capture term dependencies. More recent studies have introduced innovative approaches to information retrieval based on lexical dependency models. Gao et al. [46] proposed the Dependence Language Model, which leverages term dependencies to enhance retrieval effectiveness. Similarly, Xu et al. [47] put forth a kernel-based strategy for relevance ranking that models term dependencies in a manner akin to the schemes proposed in earlier research.

4) TOPIC MODEL

Topic modeling is another research direction that enhances information retrieval by considering semantic relationships between words. This approach typically uncovers latent text topics by modeling word associations, which allows matching queries and documents based on their topics [135]. In natural language processing tasks, topic modeling methods have become increasingly popular due to their ability to represent each dimension as a topic rather than a term. However, this can make using an inverted index impractical due to the sparsity of topic representations. Generally, topic models can be divided into probabilistic and non-probabilistic. Non-probabilistic models include latent semantic indexing [11] and non-negative matrix factorization (NMF) [136]. Wong et al. [48] proposed the Generalized

Vector Space Model (GVSM) for information retrieval, which extends the traditional vector space model by using weights to represent term frequency and importance, thereby enhancing retrieval effectiveness. Diaz [49] introduced a regularization approach for improving the accuracy of ad hoc retrieval scores.

A noteworthy probabilistic model is the Latent Dirichlet Allocation (LDA) proposed by Blei et al. [12], which captures latent topics within document collections and represents each document as a topic distribution. Experimental results reveal that the LDA-based model outperforms other state-of-the-art retrieval models. Yi and Allan [51] conducted a comparative study of various topic modeling methods, including LDA, PLSA, and LSI, and found that LDA generally performed best. Similarly, Lu et al. [52] compared LDA and PLSA in an empirical study, and their results also favored LDA for most tasks. Atreya and Elkan [53] demonstrated the limitations of Latent Semantic Indexing (LSI) for TREC collections and proposed a new retrieval model that utilizes word co-occurrence statistics to estimate document similarity. Their experimental results indicate that this method outperforms LSI and several other retrieval models.

5) MULTILINGUAL RETRIEVAL MODEL

The challenge of vocabulary mismatch in information retrieval has been tackled using various approaches, including the statistical translation method. This method extends the document representation function from merely considering frequency to incorporating translation models. This framework treats queries and documents as texts in different languages, and statistical machine translation (SMT) techniques are employed to model their relationship. Retrieval with translation models requires learning translation probabilities from queries to corresponding relevant documents, which can be obtained from labeled data, making it a supervised learning technique.

Berger and Lafferty [138] introduced the idea of formulating retrieval tasks as an SMT problem, where a query q is translated to document d with the conditional probability $P(d|q)$. The model can be expressed as:

$$P(d|q) \propto P(q|d)P(d) \tag{2}$$

In this equation, $P(q|d)$ denotes a translation model that translates d to q, while $P(d)$ represents a language model that generates d. Translation probabilities can be calculated using queries and their relevant documents, such as click-through datasets, while the language model can be trained through various methods like BM25. Karimzadehgan and Zhai [55] observed that the translation probability $P(q|d)$ allows incorporating semantic relationships among terms with non-zero probabilities, providing a form of ''semantic smoothing'' for $P(d|q)$. A crucial difference between machine translation and conventional translation for retrieval is that queries and documents are actually in the same language. The probability of translating a word to itself should be relatively high (i.e., $P(w|w) > 0$), which corresponds to exact term matching in retrieval tasks.

Further studies have offered theoretical analyses of the translation language model for information retrieval. Karimzadehgan and Zhai [57] investigated the model's properties and constraints using axiomatic analysis. Riezler and Liu [58] proposed a query expansion technique that utilizes monolingual statistical machine translation (SMT) to improve retrieval efficiency. Gao and Nie [59] introduced a query expansion method that leverages translation models and search logs to enhance retrieval effectiveness.

B. SPARSE RETRIEVAL METHODS

The growing popularity of sparse retrieval methods can be attributed to their ability to represent individual documents and queries using sparse vectors, which only activate a small number of dimensions. This approach aligns with human cognitive processes and can be easily integrated into existing indexing mechanisms, optimizing retrieval performance. Sparse retrieval approaches can be divided into two main categories. The first category involves using neural models to enhance term weighting schemes while preserving the symbolic encoding of documents and queries. This method is commonly known as neural weighting schemes. On the other hand, the second category focuses on directly obtaining sparse representations of documents and queries in the latent space through the application of neural networks. This particular technique is referred to as sparse representation learning.

1) NEURAL WEIGHTING SCHEME

There are two methods to leverage neural models in sparse term-based retrieval. The first method involves designing neural models that predict term weights based on semantics rather than relying on predefined heuristic functions. This approach allows for the re-weighting of term significance before indexing. The second strategy involves expanding each document with additional terms and indexing the expanded documents using classical term-based methods. One of the early models to learn term weights is DeepTR [60], which employs neural word embeddings to determine term importance. It constructs a feature vector for each query term and learns a regression model to map feature vectors onto ground truth term weights. These estimated weights can replace traditional term weighting schemes, such as BM25 and LM, enhancing retrieval performance for bag-of-words query representations. Another approach involves integrating neural word embeddings into information retrieval systems [61]. The authors propose a method to evaluate the effectiveness of various neural word embeddings in information retrieval and describe how to incorporate these embeddings and re-weighting into retrieval models. Context-aware term weighting methods, like ''DeepCT'' [62], improve the first-stage retrieval process's effectiveness by assigning higher weights to more relevant terms based on

their context. The authors suggest a deep learning-based method that learns to estimate term importance in a sentence or passage according to context. Frej et al. [63] presented an alternative approach to learning term discrimination in information retrieval, using a deep learning-based method that distinguishes relevant terms from irrelevant ones based on context. Expanding on "DeepCT," the authors [62] consider term importance at multiple granularity levels, from document to sentence level, enhancing retrieval effectiveness. The efficiency of the "DeepCT" approach is evaluated and compared with traditional term weighting methods in terms of computational cost and retrieval performance [65]. Lin and Ma [66] offered a conceptual framework for analyzing and comparing various information retrieval techniques, particularly neural-based methods. They introduced "DeepImpact" and "COIL" as two different framework dimensions, proposing "uniCOIL" as a unified approach that combines the strengths of both dimensions. Nogueira et al. [67] proposed a model called "Doc2Query," which predicts relevant documents given a document using a neural network. They demonstrate that this approach can improve document retrieval effectiveness. Nogueira et al. [68] enhanced their previous work by incorporating additional information, such as document titles and clicked snippets, into the Doc2Query model. They also propose a new evaluation metric, "Top-k-TTTTT," showing that the new model, "DocTTTTTQuery," outperforms the previous model on the benchmark dataset.

Mao et al. [69] proposed a model called "GAR," which generates new queries based on a given question, then retrieves relevant documents using those queries. This approach is shown to improve open-domain question-answering effectiveness. Yan et al. [70] proposed a model called "UED" that is trained on a large text corpus using a combination of supervised and unsupervised learning, effectively ranking and expanding passages. The authors [67], [68], [69], [70] proposed novel methods for expanding a given document or query by predicting or generating relevant queries and then using those queries to retrieve additional information. These methods have improved the effectiveness of document retrieval, open-domain question answering, and passage ranking. As new approaches are introduced, the field of information retrieval continues to evolve and improve. Models like MacAvaney et al.'s [71] used a neural network to predict term importance in a document and retrieve additional information. Others, such as SparTerm [72], used a neural network to learn a sparse term-based representation of a document. Likewise, SPLADE and SPLADE v2 [73] employ neural networks to learn a sparse lexical representation of a document and expand the query. DeepImpact, proposed by Mallia et al. [74], utilizes a neural network to learn the impact of passages on retrieval. TILDEv2, suggested by Zhuang and Zuccon [75], applies a neural network to learn term-based representations of passages and expand the query. Similarly, SpaDE, proposed by Choi et al. [76], uses a neural network to learn a sparse representation of a document and encode the document with two encoders. These models demonstrate the potential of neural networks in enhancing information retrieval by refining term weighting, expanding queries, or learning sparse representations.

2) SPARSE REPRESENTATION LEARNING

One key advantage of sparse representations is their ability to capture semantic relationships between words and phrases, going beyond simple term frequency measures. This allows for better handling of synonymy and polysemy issues, which can be challenging for traditional term weighting schemes. Sparse representations can also encode higher-level semantic information, enabling the retrieval system to understand better the context and meaning of a document or query. Unlike term-weighting approaches in symbolic space, sparse representation learning techniques concentrate on constructing sparse vectors for both queries and documents. These representations aim to encapsulate the semantic essence of each input text, thereby placing queries and documents within a latent space. However, unlike the topic models discussed in Section III-A(4), the dimensions of the latent space created by neural models lack distinct concepts. The resulting sparse representations can be effectively stored and searched using an inverted index. Each entry in the index table corresponds to a "latent word" rather than a conventional term. Salakhutdinov and Hinton [77] proposed a method called semantic hashes, which use a neural network to map documents to compact binary codes that can be used for efficient approximate nearest neighbor search. They showed that this approach could achieve state-of-the-art performance. Zamani et al. [78] proposed a model called SNRM, which uses a neural network to learn a sparse representation of a document and showed that it could improve the effectiveness of information retrieval. Jang et al. [79] proposed a model called UHD-BERT, which uses a neural network to learn ultra-high-dimensional sparse representations of a document and showed that it could improve full-ranking effectiveness. Yamada et al. [80] proposed a model called BPR, which uses a neural network to learn a semantic hash for each passage and showed that it could improve the efficiency of open-domain question answering. Lassance et al. [81] proposed a model called CCSA, which uses a neural network to learn a composite code sparse representation of a document and showed that it could improve the effectiveness of first-stage retrieval. All these papers propose new methods that use sparse representations to improve the effectiveness and efficiency of information retrieval systems.

C. DENSE RETRIEVAL METHODS

The advent of deep learning techniques has significantly transformed the information retrieval landscape. As seen in fig. (4), dual-encoder architecture, also known as a siamese network [29], is the typical design for dense retrieval models. It comprises twin networks that receive different inputs (queries and documents) and independently develop standalone dense embeddings for them. This section provides

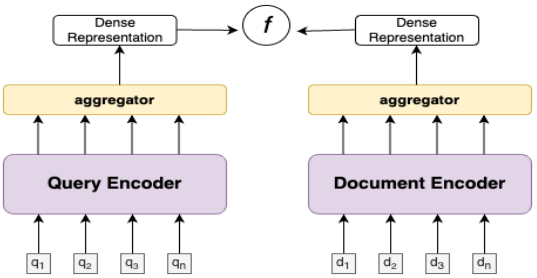

FIGURE 4. Dual-encoder architecture of dense retrieval methods.

a comprehensive overview of state-of-the-art deep learning methods for semantic information retrieval, outlining key advancements and their respective contributions to the field.

1) WORD-EMBEDDING-BASED APPROACHES

Word embeddings, which are dense and continuous word representations that capture semantic meaning, have gained popularity for document and query representation in information retrieval systems. Various techniques have been proposed to use word embeddings in combination with other methods. These include aggregation techniques [82], bilingual word embeddings [83], dual embedding spaces [85], and context representation for natural language generation [86]. In addition, researchers have turned to neural networks to map documents and queries into continuous spaces, using similarity measures for document ranking [87], [88], [89]. By incorporating these diverse approaches, information retrieval systems can harness the power of word embeddings to improve their overall performance and effectiveness.

2) TRANSFORMER-BASED APPROACHES

The advent of transformer based models have paved the way for developing various encoding techniques for questions and documents. Nie et al. [93] proposed a decoupled encoding approach using DC-BERT, which enhances document retrieval effectiveness. Similarly, Yang et al. [94] introduced a question-answering method that employs a neural retrieval component, a cross-attention mechanism for response generation, and a data augmentation technique to boost performance.

3) APPROXIMATE NEAREST NEIGHBOR SEARCH AND NEGATIVE CONTRASTIVE LEARNING APPROACHES

Recent research has also delved into the use of approximate nearest neighbor search and negative contrastive learning for dense text retrieval. Xiong et al. [95] put forth a method called ANCE, which utilizes a neural network to encode documents and queries. This is followed by applying approximate nearest neighbor search and negative contrastive learning for document ranking. Zhan et al. [96] presented an efficient technique for training dense retrieval models using a combination of hard and soft negative sampling. Meanwhile, Shan et al. [97] employed a global weighted self-attention network for web search. Other researchers have explored optimization techniques and innovative methodologies to enhance the performance of dense retrieval models in various scenarios [98], [99], [100], [101], [102], [103].

4) PASSAGE-CENTRIC APPROACHES

In an effort to further improve dense retrieval models, researchers have explored new approaches, such as passage-centric similarity relations, which concentrate on the relationship between passages rather than individual words [104]. Moreover, Khattab et al. [105] proposed relevance-guided supervision for training ColBERT, a pre-trained transformer model designed for OpenQA tasks, to enhance passage relevance learning. Singh et al. [106] introduced an end-to-end training method for the multi-document reader and retriever systems in open-domain question answering. These advancements showcase the ongoing efforts to refine and expand upon existing retrieval methods to achieve better performance in information retrieval tasks.

5) PSEUDO RELEVANCE FEEDBACK APPROACHES

Pseudo Relevance Feedback (PRF) is a widely-used technique in information retrieval that aims to improve the effectiveness of the initial query by leveraging the information obtained from top-ranked documents in the preliminary search results. The assumption behind PRF is that these documents are likely to be relevant to the user's information needs. The system can refine the query by analyzing their content and producing more accurate retrieval results. Building upon the concept of PRF, Yu et al. [107] proposed a method to enhance query representations for dense retrieval using pseudo-relevance feedback. This technique extracts relevant information from a set of retrieved documents to improve retrieval performance. Building on this, Wang et al. [108] introduced a method for training dense retrieval models utilizing pseudo-relevance feedback and multiple representations, allowing the model to learn more robust query representations.

6) DISCRIMINATIVE SEMANTIC RANKING APPROACHES IN DENSE RETRIEVAL

As researchers strive to improve information retrieval, they have developed methods to train models that can effectively distinguish between relevant and non-relevant documents in dense retrieval settings. Cai et al. [109] suggested a method for training a discriminative semantic ranker for question retrieval, focusing on this crucial aspect of accurate retrieval. To further refine the understanding of the relationship between passages and queries, Wu et al. [110] proposed a representation decoupling method that improves open-domain passage retrieval by separating the encoding of passages and queries.

Continuing this research line, Ren et al. [111] introduced a dense passage retrieval and re-ranking model training approach. Meanwhile, Lindgren et al. [112] presented a more efficient method for training retrieval models through negative caching, Lu et al. [113] proposed a technique for training

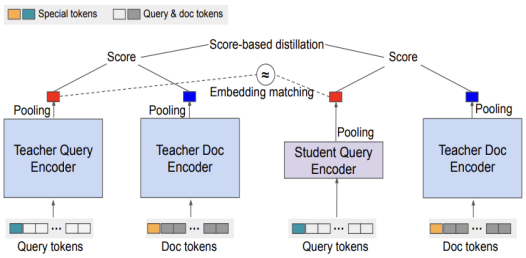

FIGURE 5. Cross-Architecture Knowledge Distillation [159].

neural passage retrieval models using multi-stage training and improved negative contrast.

7) KNOWLEDGE DISTILLATION APPROACHES

Knowledge distillation, a technique for transferring knowledge from a pre-trained, larger model to a smaller model, has been employed to enhance performance on various tasks such as document ranking [114], chat-bot systems [115], question answering [116]. Large-scale retrieval tasks [117]. Researchers have explored using pre-trained BERT models, transferring knowledge across different model architectures, and applying Margin-MSE loss functions [118] for knowledge distillation.

Kim et al. presented EmbedDistill Figure (5), an innovative distillation approach for information retrieval that leverages relative geometry among queries and documents. It improves upon traditional methods by using embedding matching for stronger local geometry signals and query generation for better global data manifold coverage. Applicable to both dual-encoder and cross-encoder models, EmbedDistill shows promising results on benchmarks like MSMARCO and NQ. The paper's theoretical analysis also supports the effectiveness of the proposed approach.

Furthermore, the relationship between pre-trained models and the effects of distilling knowledge from one model to another have also been investigated [119], [120].

8) CROSS-MODAL APPROACHES

Cross-modal techniques for dense text retrieval have gained interest as a means to bridge the gap between different modalities, such as text and images. Researchers have focused on developing methods that enable the encoding of textual and visual information into shared latent spaces for retrieval tasks [139], [140], [141]. These methods often employ deep learning techniques such as convolutional neural networks (CNNs) for image feature extraction, combined with word embeddings or recurrent neural networks (RNNs) for textual data representation. Some notable advancements in this area include using attention mechanisms [142], [143] and incorporating transformer-based models, such as BERT, for cross-modal tasks [144], [145].

9) REINFORCEMENT LEARNING APPROACHES

Reinforcement learning techniques have been employed in dense text retrieval to optimize retrieval policies and explore the interaction between users and retrieval systems [146], [147]. Reinforcement learning approaches, such as Deep Q-Networks (DQNs) and policy gradient methods, have been applied to tasks such as document ranking [148] and query auto-completion [149]. These approaches focus on learning optimal actions and strategies in response to user interactions and feedback to improve retrieval performance. Some research has also explored incorporating reinforcement learning into pre-trained transformer models, such as BERT, to fine-tune the models for specific retrieval tasks [150].

10) GRAPH-BASED APPROACHES

Graph-based methods for dense text retrieval exploit the relationships between documents, terms, and other entities to improve retrieval performance [151], [152]. Recently, graph neural networks (GNNs) have been used to model such relationships in various retrieval tasks, including document ranking [153] and question answering [154]. Graph-based approaches can leverage local and global information within the graph structure, enabling the model to capture complex relationships and dependencies between various entities [155], [156]. Moreover, integrating GNNs with pre-trained transformer models, such as BERT, has been explored to enhance the performance of retrieval tasks [157], [158].

D. HYBRID RETRIEVAL METHODS

Hybrid retrieval methods aim to improve the performance of text retrieval systems by combining different representations, architectures, and techniques. This section discusses various approaches and their contributions to the field of information retrieval. Vulić and Moens [83] presented a method that linearly combines monolingual and cross-lingual word embeddings to enhance retrieval performance. Their approach leverages the strengths of both monolingual and cross-lingual embeddings, facilitating better alignment between different languages and leading to improved results in multilingual retrieval tasks. Ganguly et al. [121] introduced the Generalized Language Model (GLM), which utilizes word embeddings to improve retrieval performance. By incorporating word embeddings into the language modeling framework, GLM captures the semantic relationships between words, allowing for better query-document matching and, thus, improved retrieval performance. Roy et al. [122] suggested a method for combining word embeddings using set operations to enhance retrieval performance. Their approach captures the semantic similarity between the query and document terms by performing set operations on word embeddings, improving retrieval performance while maintaining computational efficiency. Mitra et al. [85] proposed the dual embedding space model (DESM), which combines word embeddings to improve retrieval performance. DESM leverages two different embedding spaces to capture local and global semantic information, providing a more comprehensive representation of terms and

TABLE 2. Overview of First Stage Retrieval Techniques in Research Studies: Listing the reference, publication year, paper category, and algorithm used for first stage retrieval.

	Reference	Year	Category	Technique/Algorithms
Sparse Retrieval Methods	Zheng et al. [60]	2015	Neural Weighting Schemes	Distributed representations of words.
	Zuccon et al. [61]	2015		CBOW,SkipGram,Translation Model.
	Dai et al. [64]	2019		Deep Contextualized Term Weighting, BERT, DeepCT-Index, DeepCT-Query, BM25, Query Likelihood.
	Mitra et al. [279]	2019		BERT, Duet, CKNRM, query term independence assumption.
	Nogueira et al. [75]	2019		Document expansion, sequence-to-sequence model, neural networks, re-ranking.
	Nogueira et al. [68]	2019		Document expansion, T5, sequence-to-sequence model, BM25, top-k sampling
	Dai et al. [62]	2020		Deep Contextualized Term Weighting (DeepCT), BERT, DeepCT-Index, BM25.
	Dai et al. [256]	2020		BOW, Inverted Index Content-based weak-supervision strategy, PRF-based weak-supervision strategy.
	Frej et al. [63]	2020		Shallow neural networks, Term Discrimination Values), TF-IDF, BM25
	Mitra et al. [280]	2020		Transformer-Kernel model, query term independence assumption, Conformer layer.
	Mao et al. [69]	2020		Generation Augmented Retrieval, text generation, sparse representations, dense retrieval methods.
	Bai et al. [72]	2020		Gating Controller, unifying term-weighting
	Yan et al. [70]	2021		Generation-Augmented Retrieval, Document Expansion using Seq2Seq model.
	Mallia et al. [74]	2021		DeepImpact,DocT5Query.
	Ma et al. [281]	2022		Frequency-based weighted averaging, context-based word weighting, hybrid weighting schemes.
	SAHİN et al. [282]	2022		Pairwise ranking loss, BERT regression model, term recall, BM25-based index
	Kleinberg et al. [283]	2000	Sparse RL	Network navigation, decentralized search algorithms, local information.
	Salakhutdinov et al. [77]	2009		Deep generative model, autoencoder, backpropagation, noise injection.
	Zamani et al. [78]	2018		Sparsity introduction, L1-norm minimization.
	Karpukhin et al. [88]	2020		Dual-encoder approach.
	Zoph et al. [284]	2022		Mixture-of-Experts (MoE), Switch Transformers, Sparse model scaling, Fine-tuning, Transfer learning
Dense Retrieval Methods	Kenter et al. [84]	2015	Term-level RL	Neural Networks, Attention Mechanisms, Pointer Networks
	Mitra et al. [85]	2016		Dual Embedding Space Model, word2vec embedding model, cosine similarity, linear mixture, BM25.
	Seo et al. [89]	2018		Phrase-Indexed Question Answering, Document and Question Encoders, Inner Product Interaction.
	Seo et al. [125]	2019		Indexable Phrase Representations, Dense-sparse Phrase Encoding, Optimization Strategies
	Lee et al. [126]	2019		Query-agnostic indexable phrase representations, dense-sparse phrase encoding, optimization strategies.
	Feldman et al. [288]	2019		Iterative retrieval, Joint vector representation, Contextualized sentence-level representations.
	Nie et al. [93]	2020		Dual BERT models, online BERT, offline BERT, decoupled contextual encoding.
	Khattab et al. [257]	2020		BERT, late interaction architecture, fine-grained similarity modeling, vector-similarity indexes.
	Cao et al. [286]	2020		Decomposed Transformer Model, Question-Wide and Passage-Wide Self-Attentions.
	Gao et al. [285]	2021		Contextualized exact match, inverted list index, deep language models.
	MacAvaney et al. [287]	2020		Precomputing Transformer Term Representations, Compression Layer.
	Clinchant et al. [82]	2013	Document-level Representation Learning	Word Embeddings, BOW, Non-Linear Projection, Probabilistic Mixture Models, LSI.
	Huang et al. [176]	2013		Deep Structured Semantic Models, word hashing, multiple hidden-representation layers
	Hu et al. [178]	2014		Convolutional Neural Networks, layer-by-layer composition and pooling.
	Tan et al. [294]	2015		BiLSTM) models, cosine similarity, CNN, attention mechanism.
	Ai et al. [289]	2016		Paragraph Vector Model, Negative Sampling, Weighting Scheme, Vector Norms.
	Liu et al. [292]	2016		Constrained Word Embeddings, Word Embedding Weighting.
	Henderson et al. [86]	2017		n-gram embedding features, dot-product optimization, efficient search algorithm.
	Gillick et al. [87]	2018		Dual Encoders, Approximate Nearest Neighbor (ANN) Search, Negative Sampling.
	Gysel et al. [290]	2018		Neural Vector Space Model, Gradient Descent, Latent Semantic Embedding, Lexical Language Models.
	Lee et al. [237]	2019		Joint learning of retriever and reader, pre-training with Inverse Cloze Task.
	Tamine et al. [293]	2019		Online/Offline Learning Approaches, Knowledge Distillation, Regularization, Relational Constraints.
	Humeau et al. [296]	2019		Pre-trained transformers, Poly-encoders, Bi-encoders, Cross-encoders.
	Agosti et al. [291]	2020		Knowledge-enhanced neural models, Multi-task learning, Word and concept representations
	Lin et al. [114]	2020		ColBERT, Knowledge Distillation, MaxSim Operator, Dot Product, ANN Search, Document Expansion.
	Karpukhin et al. [88]	2020		Dense representations, dual-encoder framework.
	Zhan et al. [91]	2020		RepBERT, contextualized embeddings, inner product scoring, first-stage retrieval.
	Vakili et al. [115]	2020		Cross-encoder architecture, Bi-encoder architecture, Knowledge distillation.
	Guu et al. [239]	2020		Masked language modeling, attention mechanisms.
	Chang et al. [238]	2020		Transformer models, embedding-based retrieval models, pre-training tasks (ICT, BFS, WLP), BM-25.
	Liang et al. [295]	2020		Pre-training with latent knowledge retriever, masked language modeling,, fine-tuning.
	Luan et al. [297]	2021		Dual encoders, attentional neural networks, sparse-dense hybrids.
	Tang et al. [298]	2021		Dense representations, Bi-encoder, iterative clustering, approximate nearest neighbor search library.
Hybrid Retrieval Methods	Vulić et al. [83]	2015	Hybrid Retrieval Methods	Bilingual Word Embeddings Skip-Gram, ad-hoc retrieval, pseudo-relevance feedback modeling.
	Ganguly et al. [121]	2015		Word embeddings, Bilingual Word Embeddings Skip-Gram model, document embeddings, ad-hoc MoIR.
	Dos et al. [124]	2015		Bag-of-words, Convolutional Neural Network.
	Mitra et al. [85]	2016		Word embeddings, Dual Embedding Space Model (DESM), cosine similarity.
	Roy et al. [122]	2016		Word vector embeddings, similarity metric, set-based representation.
	Seo et al. [125]	2019		Dense-sparse phrase encoding, optimization strategies.
	Lee et al. [126]	2019		Contextualized Sparse Representation (SPARC), Phrase Retrieval Model, Rectified Self-Attention.
	Kuzi et al. [128]	2020		Deep Neural Networks, Lexical Models, Hybrid Approach.
	MacAvaney et al. [71]	2020		Bag-of-words, Convolutional Neural Network.
	Luan et al. [297]	2021		Dual encoders, attentional neural networks, sparse-dense hybrids.
	Gao et al. [127]	2021		Residual-based embedding learning, neural embedding matching model, error-based negative sampling.
	Zhu et al. [299]	2022		Pseudo-relevance feedback, Loss-over-Loss framework.

leading to better retrieval performance. Combining local and global embedding spaces allows the model to capture nuances and relationships between words, ultimately resulting in improved query-document matching. Boytsov et al. [123]

introduced a method that replaces term-based retrieval with k-NN search while incorporating translation models and BM25 to improve retrieval performance. This approach enables the model to consider the semantic relationships between terms and the traditional statistical weighting schemes, resulting in a more effective retrieval system. Dos Santos et al. [124] proposed a method that combines Bag-of-Words (BOW) and Convolutional Neural Networks (CNN) to enhance question-answering performance. By integrating the strengths of BOW, which captures term frequency information, and CNN, which captures local semantic relationships between words, their approach achieves a more comprehensive representation of text and improved performance in question-answering tasks. Seo et al. [125] introduced DenSPI (Dense-Sparse Phrase Index), a method designed to improve real-time question-answering performance. DenSPI combines dense and sparse representations to capture fine-grained and coarse-grained semantic information, enabling efficient and accurate retrieval of relevant passages for question answering. Lee et al. [126] proposed SPARC (Sparse, Contextualized Representations) to enhance real-time question-answering performance. SPARC leverages contextualized representations to encode the interactions between terms within a text and capture the context-specific meanings of words. By combining these contextualized representations with sparse term-based features, SPARC provides a richer text representation, leading to improved question-answering performance. Wrzalik and Krechel [92] introduced CoRT (Complementary Rankings from Transformers), combining transformer-based models with traditional retrieval methods such as BM25 to improve retrieval performance. CoRT leverages the strengths of both deep learning-based models and traditional ranking algorithms to create an ensemble system that achieves better retrieval performance than either method alone. Gao et al. [127] introduced CLEAR (Complement Lexical Retrieval Model), which combines lexical and semantic residual embeddings to improve retrieval performance. CLEAR leverages the complementary nature of lexical and semantic information to create a more comprehensive representation of text, resulting in improved query-document matching and retrieval performance. Kuzi et al. [128] proposed a hybrid approach that combines semantic and lexical matching to improve the recall of document retrieval systems. This method enhances the retrieval system's ability to identify relevant documents by considering the semantic relationships between terms and their lexical co-occurrence patterns. This leads to improved recall and overall retrieval performance. Lin et al. [132] introduced uniCOIL (unified Conceptual framework for Information Retrieval), a conceptual framework that aims to unify various information retrieval techniques. By providing a common ground for diverse retrieval methods, uniCOIL facilitates the development and comparison of novel hybrid retrieval approaches, ultimately driving advancements in the field of information retrieval. Chen et al. [129] proposed CORW (Contextualized Offline Relevance Weighting), a method that improves the efficiency and effectiveness of neural retrieval by utilizing context and relevance weighting. CORW combines contextualized representations with relevance weighting to create a more efficient retrieval system that captures semantic nuances and relationships between terms, leading to improved retrieval performance. Arabzadeh et al. [130] introduced a method for predicting efficiency and effectiveness trade-offs for dense versus sparse retrieval strategies. This approach provides a systematic way to balance the trade-offs between computational efficiency and retrieval effectiveness, enabling the development of more practical and scalable retrieval systems. Leonhardt et al. [131] proposed Fast Forward Indexes. This method aims to improve the efficiency of document ranking by using a forward index that stores the positions of terms within documents. This approach reduces the computational overhead associated with traditional document ranking methods, enabling faster and more efficient retrieval without sacrificing the effectiveness of the ranking process. Lin et al. [132] introduced Representational Slicing, a method that densifies sparse representations for passage retrieval. By transforming sparse representations into denser ones, Representational Slicing captures more fine-grained semantic information and relationships between terms, leading to improved passage retrieval performance. In general, hybrid retrieval methods have demonstrated the potential to improve text retrieval system performance by combining different representations, architectures, and techniques. These methods take advantage of the strengths of diverse approaches, such as word embeddings, contextualized representations, attention mechanisms, and traditional ranking algorithms, to create more effective and efficient retrieval systems. As research in this area continues to advance, it is expected that novel hybrid retrieval methods will further enhance the performance of information retrieval systems, enabling users to find relevant information more quickly and accurately.

IV. SECOND STAGE - RANKER

In the modern era of information retrieval and web search, the ranking has become essential to provide users with relevant and high-quality search results. As a critical component of the search engine pipeline, the second stage ranker refines the ranking of the initially retrieved documents to improve the quality of search results. This section delves into various learning-to-rank techniques and deep learning-based ranking models that enhance ranking tasks' performance.

A. LEARNING TO RANK

Techniques for term weighting, like BM25, are typically categorized as unsupervised methods, even though they possess adjustable parameters that can be tweaked using learning data [18]. Significant progress within text ranking started end of the 1980s with the introduction of supervised machine learning algorithms to create ranking models, with early examples being the work of [160], [161], and [162]. This method, known as ''learning to rank'' (LTR), is heavily dependent on manually crafted features, focusing mainly on

TABLE 3. Overview of Second Stage Ranking Techniques in Research Studies: Listing the reference, publication year, paper category, and algorithm used for second stage-ranker.

	Reference	Year	Category	Technique/Algorithms
Learning To Rank	Fuhr et al. [160]	1989	Learning To Rank	Polynomial functions, probabilistic models, description vectors.
	Gey et al. [161]	1994		Logistic Regression, Standardization.
	F.C Gey et al. [162]	1993		Adaptive Bilinear Retrieval Model, Feedforward Neural Network, Perception Learning Algorithm.
	Joachims et al. [165]	2002		Support Vector Machines (SVM).
	Radlinski et al. [166]	2005		Learning to Rank, Query Chains, Clickthrough Data, Ranking SVM.
	Pasumarthi et al. [171]	2019		Deep Learning, TensorFlow.
	Jingtao et al. [101]	2021		A query-side training Algorithm for Directly Optimizing Ranking pErformance (ADORE).
	Oosterhuis et al. [300]	2022		Plackett-Luce gradient estimation, sampling techniques
	Jia et al. [301]	2022		Representation learning, RankNet, LambdaRank, Neural tangent kernel.
	Wu et al. [302]	2022		Deep Learning, Peer Learning.
Deep Learning Ranking Model	Bromley et al. [175]	1993	Representation Based Models	Siamese Neural Network.
	Schuster et al. [183]	1997		Bidirectional Recurrent Neural Network (BRNN).
	Huang et al. [176]	2013		Deep Structured Semantic Models (DSSM), word hashing, clickthrough data optimization.
	Socher et al. [185]	2013		Neural Tensor Networks, Word Vectors.
	Kalchbrenner et al. [186]	2014		Convolutional Neural Network, Dynamic k-Max Pooling.
	Shen et al. [177]	2014		Convolutional Neural Networks, Latent Semantic Models, Convolution-Max Pooling Operation.
	Hu et al. [178]	2014		Convolutional Neural Networks.
	Shen et al. [180]	2014		Convolutional Neural Network, Latent Semantic Analysis.
	Qiu et al. [179]	2015		Convolutional Neural Networks, Tensor Layer.
	Mueller et al. [181]	2016		Long Short-Term Memory (LSTM) network, word embedding vectors.
	Wan et al. [184]	2016		Bidirectional Long Short-Term Memory (Bi-LSTM), k-Max Pooling, Multi-Layer Perceptron.
	Nalisnick et al. [186]	2016		Word2vec, Dual Embedding Space Model, Cosine Similarity.
	Mitra et al. [85]	2016		Word2Vec, Dual Embedding Space Model, cosine similarity, linear mixture model.
	Guo et al. [21]	2016	Interaction Models	Matching Histogram Mapping, Feed Forward Matching Network, Term Gating Network.
	Yang et al. [191]	2016		Attention-based Neural Matching, Value-Shared Weighting Scheme, Question Attention Network.
	Pang et al. [192]	2016		MatchPyramid Model, Convolutional Neural Network, Interaction Function, Pooling Size, Kernel Size.
	Pang et al. [193]	2016		DeepRank, CNN, 2D-GRU, query term importance, term gating mechanism.
	He et al. [194]	2016		Pairwise Word Interaction Model, Similarity Focus Layer, ConvNet, BiLSTMs.
	Wan et al. [197]	2016		Recursive matching, Spatial RNN, Gates.
	Hau et al. [198]	2017		Position-Aware Convolutional Recurrent Relevance Matching Model.
	Jaech et al. [195]	2017		, Match-Tensor architecture, Canonical Correlation Analysis, Recurrent Neural Networks.
	Xiong et al. [189]	2017		Kernel-based neural model, Translation matrix, Kernel-pooling technique, Learning-to-rank layer.
	Fan et al. [190]	2017		Hierarchical Neural Matching, Local Matching Layer, Global Decision Layer, Deep Neural Networks.
	Dai et al. [200]	2018		Convolutional Neural Networks, Kernel Pooling, Learning-to-Rank.
	Lan et al. [201]	2018		Subword-based Models, Character and Character n-gram representations, Multi-task Learning.
	Tang et al. [203]	2019		Neural Information Retrieval, DeepTileBars Model, Topical Segments, Interaction Matrix.
	Abri et al. [196]	2022		LSTM, topic modeling, re-ranking.
	Mitra et al. [22]	2017	Hybrid	Deep Neural Networks, Local Representations, Distributed Representations.
	Nie et al. [202]	2018		Convolutional Neural Networks (CNN), Interaction-based Models, Multi-level Matching Models.
	Mitra et al. [204]	2019		Duet v2 model.
	Ruiyang et al. [111]	2021		Unified Training Approach, Hybrid data augmentation strategy.
	Yin et al. [211]	2016	Attention	Attention Based Convolutional Neural Networks (ABCNN), three attention schemes.
	McDonald et al. [199]	2018		Context-Sensitive Neural, Convolutional Neural Networks.
	Tan et al. [209]	2018		Multiway Attention Networks, Matching-aggregation Framework, Word Embeddings.
	Kim et al. [207]	2019		Densely-connected Co-attentive Recurrent Neural Network (DRCN), DenseNet, Autoencoder.
	Zhang et al. [208]	2019		Attention mechanism, Dynamic Re-read (DRr) unit, Attention Stack-GRU (ASG) unit.
	Zhang et al. [303]	2021		Locally Aware Dynamic Reread Attention Network (LadRa-Net).
	Zhang et al. [304]	2022		Attention mechanism, Semantic matching.
	Li et al. [220]	2018	Pre Transformers	Pseudo relevance feedback, end-to-end neural framework.
	Yang et al. [212]	2019		BERT, sentence-level inference, score aggregation.
	Yangb et al. [214]	2019		Generalized Autoregressive Pretraining, Permutation Language Modeling, Transformer-XL.
	Dai et al. [215]	2019		Transformer-XL, segment-level recurrence mechanism, novel positional encoding scheme.
	Dai et al. [216]	2019		BERT, contextual neural language model, domain adaptation.
	Nogueira et al. [218]	2019		BERT, monoBERT, duoBERT, multistage ranking architecture.
	Nogueira et al. [221]	2019		Pretrained language models (ELMo, OpenAI GPT, BERT).
	Boualili et al. [219]	2020		BERT, Exact Match Signals, Marking Technique.
	Zhang et al. [303]	2022		Attention mechanism, BERT, Semantic matching.

the statistical attributes of terms within texts and the inherent qualities of the texts themselves.

Statistical attributes of terms include document frequencies, document lengths, term frequencies, and other elements present in scoring functions like BM25. As a matter of fact, BM25 scores and other precise matches scoring functions are frequently used as features within a learning-to-rank framework, with features occasionally integrating field-specific proximity constraints [163]. Inherent qualities of texts vary from basic statistics; in the context of web searches, hyperlink graph features, including inbound and outbound link counts and PageRank scores, are also prevalent [164].

Real-world search engines may utilize hundreds or even more features [163]. For systems with large user bases,

user behavior-based features, such as query frequency or link click frequency in various contexts, serve as significant importance indicators and are fully merged into learning-to-rank techniques.

The rise of learning to rank was chiefly spurred by the escalating prominence of search engines as vital mechanisms for browsing the internet, as earlier techniques reliant on human-curated directories turned unfeasible owing to the rapid increase of obtainable content. Log data, which records user actions such as inquiries and clicks, can be harnessed to refine machine-learned ranking frameworks [165], [166].

Upgraded search experiences prompted a larger user base, yielding more records data and conduct-focused aspects to refine ranking quality further. Learning-to-rank techniques can be broadly divided into three categories based on their loss functions' general forms: pointwise, pairwise, and listwise approaches [167], [168].

While this classification mainly concentrates on loss function forms, it can additionally be utilized to depict ranking methods using transformers. The zenith of learning to rank took place at the start of the decade, just before the deep learning evolution, within the creation of tree ensemble-based models, specifically gradient-boosted decision trees [169], [170].

Although transformers for text ranking are additionally seen as a supervised machine-learning technique, they are not typically considered learning-to-rank methods. Learning to rank is distinguished by its multiple sparse-hand-engineered features, as opposed to the deep learning approaches that succeeded it [19], [161].

Nevertheless, the term ''deep learning to rank'' shows arisen in recent discussions to represent deep learning methods that also integrate sparse features [171].

Transformers have transformed the area of natural language processing and also have been effectively employed in various tasks, including text ranking. In contrast to learning-to-rank methods, which typically use hand-crafted features, transformer-based approaches harness deep learning to learn intricate representations of input texts and produce rankings. This transition has facilitated the development of more advanced text ranking models capable of better capturing semantic relationships and context, ultimately leading to improved search engine performance and user experiences.

B. DEEP LEARNING BASED RANKING MODEL

Following the learning-to-rank era, deep learning emerged as the next significant development within text ranking, initially gaining traction in computer vision and subsequently in natural language processing communities. Deep learning approaches were intriguing from the information retrieval perspective due to two main factors. First, continuous vector representations allowed text retrieval to surpass the limitations of exact term matching. Second, neural networks eliminated the requirement for labor-intensive manually created features, which was a significant challenge in constructing learning-to-rank systems. Within the realm of DL methods for text ranking, it is helpful to differentiate between pre-BERT models and BERT-built models, as the BERT revolution was a driving force behind the advancements in the field.

The Deep Learning Track at TREC 2019, the initial extensive assessment of retrieval methods after the debut of BERT, demonstrated the influence of pre-trained neural language models on retrieval efficiency among several team's approaches [172].

These pre-BERT models explored different neural architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants. These models aimed to capture semantic information and context within the text to enhance the significance of document ranking. While some models leveraged supervised learning, others utilized unsupervised or semi-supervised strategies to learn representations of text data [14], [173], [174].

However, the opening of BERT and transformer-based models significantly impacted the field of text ranking. These models, pre-trained on considerable portions of text data, demonstrated unparalleled performance across various natural language processing tasks, including document ranking. By learning rich contextual representations and overcoming the limitations of previous deep learning models, transformer-based approaches have set new standards for effectiveness in text ranking. These studies exhibit considerable architectural similarities by excluding another extensive body of literature, primarily from the NLP society, focusing on the near corresponding situation of determining semantic likeness between two sentence models. In this regard, ideas are exchanged between the IR and NLP communities.

Nevertheless, a significant difference exists: information to a model for calculating semantic similitude exists symmetric, whereas queries and documents are different and cannot exist interchanged as model inputs. This difference implies that architectures for computing semantic likeness are typically symmetric though not necessarily for modeling query-document relevancy.

Neural ranking models can typically be divided into three categories: interaction-based, representation-based models, and hybrid representation and interaction models. Representation-based models Fig. (6) concentrate on learning dense vector representations of queries and documents alone. These can be compared using a straightforward metric like cosine likeness or inner products to determine relevance. Otherside, interaction-based models Fig. (7) approximate the representations of terms in the query and document, resulting in a likeness matrix catching term interactions. This matrix is additionally analyzed to produce a relevancy score. In both circumstances, models can employ various neural elements, such as CNN and RNN, to extract relevant signals.

Representation and interaction models are generally trained end-to-end using relevancy determinations with solely the embeddings of query and document terms serving as intake. Additional features are usually not incorporated, representing a significant departure from learning-to-rank

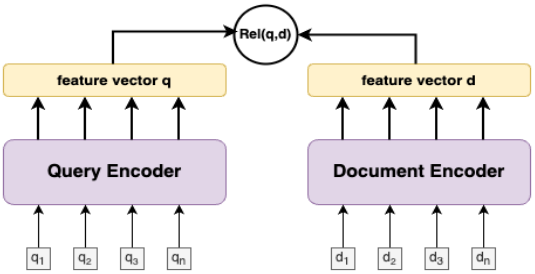

FIGURE 6. A general structure of representation-focused models is provided. These models employ two deep neural networks to transform the query and document into feature vectors. Subsequently, a ranking function Rel is utilized to convert the feature vectors of both the query and document into a relevance score expressed as a real number.

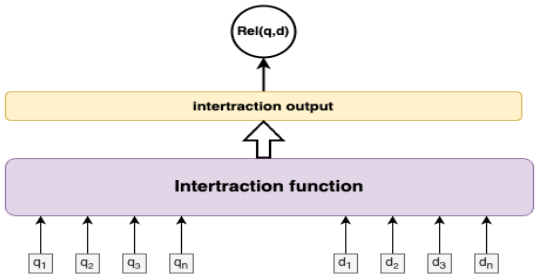

FIGURE 7. A general overview of the interaction-centric model architecture is presented. These models employ an interaction function to transform the query and the document into an interaction outcome. Subsequently, a ranking function Rel is utilized to convert this interaction output into a relevance score expressed as a real number.

techniques. The following provides more detail and examples of these models:

1) REPRESENTATION-BASED MODELS

The representation models, as illustrated in Fig. (6), utilize two independent neural network models to represent a query and a document into feature vectors query and document [175]. These models calculate the relevancy score using a query document duos by utilizing straightforward functions like cosine likeness or a Multi-Layer Perceptron among the query and document representations. Huang et al. [176] presented the foremost deep neural ranking model, the DSSM, established on the Siamese architecture. Shen et al. [177] trained Convolutional Deep Structured Semantic Model utilizing a CNN rather than feed-forward networks. ARC-I [178] also utilizes CNNs to extract feature representations of queries and documents. Qiu and Huang [179] and Shen et al. [180] developed Convolutional Neural Tensor Network and Model, respectively, incorporating CNN as the primary component.

Recurrent neural networks (RNN) have been successful in representing sentences as fixed-length feature vectors, with Manhattan LSTM (MaLSTM) by [181] and LSTM-RNN by [182] employing two LSTM models as feature extractors. Bidirectional LSTM (bi-LSTM) [183] has been used in MV-LSTM [184] to apprehend semantic matching in individual positions of the document and query by developing positional sentence representations. The model then utilizes a tensor layer [185] to model exchanges between the developed features. To extract the top k strongest interactions, k-max pooling [186] is applied in the tensor layer, followed by an MLP to calculate the relevance score.

Representation models, like the DSSM [176], understand vector representations of queries and documents to compute query–document relevancy scores. Shen et al. [180] enhanced upon DSSM by employing CNNs to apprehend context. The Dual Embedding Space Model [85], [187] illustrates texts utilizing pre-trained word2vec embeddings [188] and calculates relevancy scores by aggregating cosine likenesses across every query–document terms. Language models based on word embeddings [121] can furthermore be classified as representation models.

2) INTERACTION-BASED MODELS

Interaction-focused models Fig. (7) address the risk of missing crucial matching signals in document retrieval tasks, which is a challenge in representation-focused models. These models initiate by constructing regional interactions for a query document duo employing basic representations and then train a deep model to identify essential interaction relations among the query and document. Interaction models catch corresponding signs among the query and document them at an earlier phase.

Guo et al. [21] introduced the Deep Relevance Matching Model that employs histogram-established attributes for term matching. The interaction matrix among the query and document is calculated using pairwise cosine likenesses among query and document token embeddings. To regulate the assistance of individual query tokens to the final relevancy score, the authors suggested a term gating network with a softmax operation.

K-NRM [189] uses kernel pooling for soft-match signs to address the non-differentiability and computational inefficiency of the histogram-based representation in DRMM. Other models utilizing cosine similarity interaction matrix include the Hierarchical Neural maTching model (HiNT) [190], aNMM [191], MatchPyramid [192], and [193]. In addition to cosine similarity, similarity measures such as (.) product and indicator function are utilized in HiNT and MatchPyramid, and Gaussian Kernel is introduced in MatchPyramid [192] with considerable relations matrices.

Various architectures are employed for feature extractors to construct query-document relations and for ranking to pull corresponding signals from interactions of query and document tokens.

a: LSTM-BASED RANKING MODELS

Models like [190], [194], and [195] utilize LSTM in neural ranking models. He and Lin [194] employed bi-LSTMs for context modeling of text intakes, while [195] used two independent bi-LSTMs to apply queries and documents to hidden states. Fan et al. [190] suggested a variant of the HiNT model that sequentially gathers the signs from each passage in the document using an LSTM model and a dimension-wise k-max pooling layer. Aberi et al. [196] presented a

topic-based LSTM model to re-rank study outcomes based on a submitted input query using the previous query sequence and user click history. The model incorporates the topic distribution of user documents into the LSTM model.

b: GRU-BASED RANKING MODELS

Match-SRNN [197] uses a 2-D GRU to accumulate matching signals. Fan et al. [190] use a spatial GRU in their neural ranking model to pull relevancy corresponding evidence from query-document interaction tensors. DeepRank [193] computes a query-centric feature vector using the GRU network.

c: CNN-BASED RANKING MODELS

CNNs are employed in various interaction-focused models, including ARC-II [178], PACRR [198], PACRR-DRMM [199], Match-Tensor [195], Conv-KNRM [200] and [201], [202], [203]. ARC-II [178] is an interaction-based method that operates directly on the interaction matrix. Hui et al. (2017) proposed the PACRR model to capture position-dependent information. McDonald et al. (2018) introduced PACRR-DRMM, which adapts the PACRR model to incorporate contextual information for per query token. Jaech et al. [195] prepared Match-Tensor to explore numerous channel models for the interaction tensor. Dai et al. [200] further investigated n-gram soft matching within the Conv-KNRM model using CNN filters.

3) COMBINING REPRESENTATION AND INTERACTION IN HYBRID MODELS

By integrating them into a single model, retrieval models can bring the edge of representation and interaction-based deep architectures.

In DUET [22], a network focusing on interaction, called the local model, is merged with a representation-oriented network, named the distributed model, to form a unified deep learning architecture. The regional model processes the interaction matrix of the query and document, which is established on the exact matches of query terms within a document. This matrix is then passed through a CNN [22].

The convolutional output passes through two fully connected layers, a dropout layer and a final fully connected layer, producing a relevancy score. On the other hand, the distributed model generates a lower-dimensional feature vector for the query and document using a word embedding established representations for encoding query and document terms. After applying a sequence of nonlinear transformations to the embedded intake, the matching among query and document representations is estimated utilizing an element-wise product. The last DUET architecture computed value is the sum of local and distributed networks' scores.

Through their observed analysis, Nie et al. [202] demonstrated that interaction-based neural architectures normally outperform representation-centric architectures in information retrieval tasks. While representation-concentrating models provide the benefit of better estimation by maintaining a consistent feature vector for a document across every study. However, they miss matching signals across distinct tasks and datasets due to their static feature representation. In contrast, interaction-focused neural networks can be computationally demanding, requiring pairwise similarities between query and document token embeddings. However, they hold the edge of comprehending related cues from interacting two intakes at the initial phases.

Another reasonably available hybrid model is the DUET model [204], which supplements a representation knowledge component with an interaction part accountable for determining precise term matches. Lin [205] asked a provocative question: Are neural ranking models superior to the ''classic'' term matching approach in the lack of extensive training data obtained from behavior records? This question is crucial because academic researchers consistently struggle to access such data, typically only available to industry researchers. To what degree accomplish neural ranking models ''work'' on the restricted shares of training data that are publicly obtainable? Yang et al. [206] addressed this question by corresponding to different notable interactions and representation-based neural ranking models. Underneath these shot data conditions, almost all neural ranking methods could not perform effectively.

4) ATTENTION BASED REPRESENTATION

Attention-based models have become popular recently, and many different approaches have been proposed. These models used an attention mechanism to concentrate on relevant aspects of input text and create better representations for various NLP tasks. McDonald et al. [199] presented the Element-wise Attention-Based approach, which employs exposed-context embedding and attention importance. This technique evaluates the importance of attention for each query token relative to document tokens and generates an attention-based representation of the document based on these weights. The encoding of the query token is familiar to the document through multiplication, and the ultimate relevance score is calculated using the DRMM algorithm.

In their work, Kim et al. [207] integrated the attention mechanism into the DRCN architecture by leveraging residual connections and co-attention. This enabled the model to concentrate on pertinent tokens in the input texts. Additionally, they suggested merging feature vectors from preceding layers before computing attention weights.

The DRr-Net proposed by Zhang et al. [208] also pursued a comparable approach by implementing an attention stack-GRU unit and a Dynamic Re-read (DRr) unit. This model prioritizes significant words and operates based on attention weights in each step.

Tan et al. [209] proposed the Multiway Attention Network that leverages multiple attention procedures to improve semantic matching. The attention procedures include bilinear, concatenated, element-wise dot product, and difference of

two vectors. The model aggregates their results utilizing a bi-directional GRU network and joined attention mechanism.

Wang and Jiang [210] used multiple comparison techniques to match token embeddings and their contexts. These techniques include neural tensor networks, neural network layers, cosine similarity, Euclidean distance, and element-wise functions for vectors.

Yin et al. [211] presented an Attention Based Convolutional Neural Network that includes the attention mechanism to all input layers and the feature maps acquired from the convolutional filter. This model calculates attention significances on the intake embedding to enhance feature maps and reweights feature maps for attention-based avg pooling.

5) PRETRAINED TRANSFORMERS FOR TEXT RANKING

The advent of BERT by [5] marked a significant milestone in the domain of NLP. BERT has arisen effectively applied to various assignments, containing question-answering (QA) and document retrieval. BERT's 512 token input limit poses challenges for ad-hoc document retrieval, especially when dealing with longer documents. Yang et al. [212] proposed splitting documents into sentences and employing BERT for each one. This method, inspired by [213], suggests that single excerpts are more useful than entire documents for increased recall in retrieval. To generalize this concept, [212] introduced the top-k sentences established on BERT-calculated retrieval scores within the sentence duo classification context. To long document tasks, XLNet [214] utilizes TransformerXL [215] rather than BERT. TransformerXL incorporates close positional encoding and a component recurrence mechanism to apprehend more extended-term dependencies. Qiao et al. [216] fine-tuned BERT on two retrieval tasks, proposing four BERT-based ranking models that employ interaction and representation. BERT is more effective when working with semantically close text pairs, benefiting from relevance matching techniques. MacAvaney et al. [24] introduced a model that combines [CLS] representation with neural rankers, apprehending relevancy and semantic matching. Dai and Callan [217] enhanced BERT-based rankers using search logs, demonstrating the benefits of tuning with extensive search knowledge.

Nogueira et al. [218] proposed a multi-stage ranking architecture using a pointwise ranking strategy (monoBERT) and a pairwise learning strategy (duoBERT) to get the absolute ranked index of documents.

Boualili et al. [219] suggested incorporating exact matching signals directly in BERT's sentence classification setting for document retrieval. The successful application of BERT to the MS MARCO passage ranking [220] by [221] encouraged the research community to produce their impacts, address constraints, and expand the result differently. The availability of the MS MARCO dataset made the exploration of neural benchmarks for ranking accessible to academic research groups, as noted by [222].

V. DATA-SETS

Table (4) summarizes the state-of-the-art datasets used in the survey paper. The table includes datasets for various NLP tasks such as Passage-Retrieval [220], Bio-Medical Information Retrieval [223], [224], [225], Question Answering [226], [227], [228], [229], [230], Argument Retrieval [231], [232], Community QA [118], Entities [233] and Search Query [234]. The table lists the task, dataset name, domain, and corpus size for each dataset. The datasets are organized into sections based on the task they are used for. The table briefly describes each dataset, including its source, as mentioned in the references, and the corpus size. This table can be used as a reference for researchers to select appropriate datasets for their NLP tasks and to cite the datasets used in their research.

VI. PRESENT CHALLENGES AND FUTURE PROSPECTS

In this section, we delve into some unresolved issues and potential future developments in the context of semantic models for the initial retrieval stage. Certain aspects are critical but have not been adequately addressed in this domain, while others present exciting avenues for upcoming research.

Fig. (8) illustrates the key research areas and sub-topics discussed in the "Present Challenges and Future Prospects" section. It highlights the main theme focused on unresolved issues and potential future developments in the context of information retrieval. The diagram provides a clear overview of the critical research areas and their corresponding challenges or future directions, allowing readers to quickly understand the essential aspects that warrant further exploration in the field of IR.

A. PROGRESS IN CUSTOM PRE-TRAINING OBJECTIVES AND ARCHITECTURES FOR IR

Even though broad-spectrum pre-trained language models effectively learn global linguistic understanding, creating pre-training and fine-tuning techniques closely related to downstream tasks is a more proficient strategy for enhancing performance in specialized tasks [235], [236]. Initial exploration has occurred in pre-training objectives, model structures, and model calibration methods for IR; however, a more in-depth examination is required.

New Pre-Training Objectives: Pioneering research in pre-training objectives customization has been carried out by researchers such as Lee et al. [237], Chang et al. [238], Guu et al. [239], Ma et al. [240], [241]. For instance, Lee et al. [237] proposed the Inverse Cloze Task for retrieval tasks in a large-scale document collection. Chang et al. [238] introduced Body First Selection and Wiki Link Prediction to capture inner-page and inter-page semantic relations for passage retrieval in QA tasks. Ma et al. [240], [241] presented the Representative Words Prediction objective,

TABLE 4. SOTA Datasets in Information Retrieval. We provide the corresponding task, domain, and corpus for each dataset. For additional datasets refere [307].

Task	Data-set	Domain	Corpus
Passage-Retrieval	MS MARCO [220]	Real Queries	8.84M
Bio-Medical	CORD-19 [223]	Scholarly Articles	171K
Information	BioASQ [224]	Human-annotated QA	14.91M
Retriveval	NFCorpus [225]	Medical Information	3.6K
	Natural Questions [226]	Real Users QA	2.68M
Question	HotpotQA [227]	Question Answering QA	5.23M
Answering	FiQA-2018 [228]	Financial Opinion QA	57K
	QASC [229]	MCQ	17M
	InsuranceQA [230]	InsuranceQA	24K
	SQuAD [305]	StanfordQA	107K
	WikiQA [306]	WikiQA Corpus	3K
Argument	Arg-Microtexts Synthesis [231]	Argument	8.67K
Retrieval	Conversational Argument [232]	Conv. Argument	382K
Community QA	CQADupstack [118]	Community QA	457K
Search	MSLR-WEB10K [233]	Search Queries	10K
Query	MSLR-WEB30K [234]	Search Queries	30K

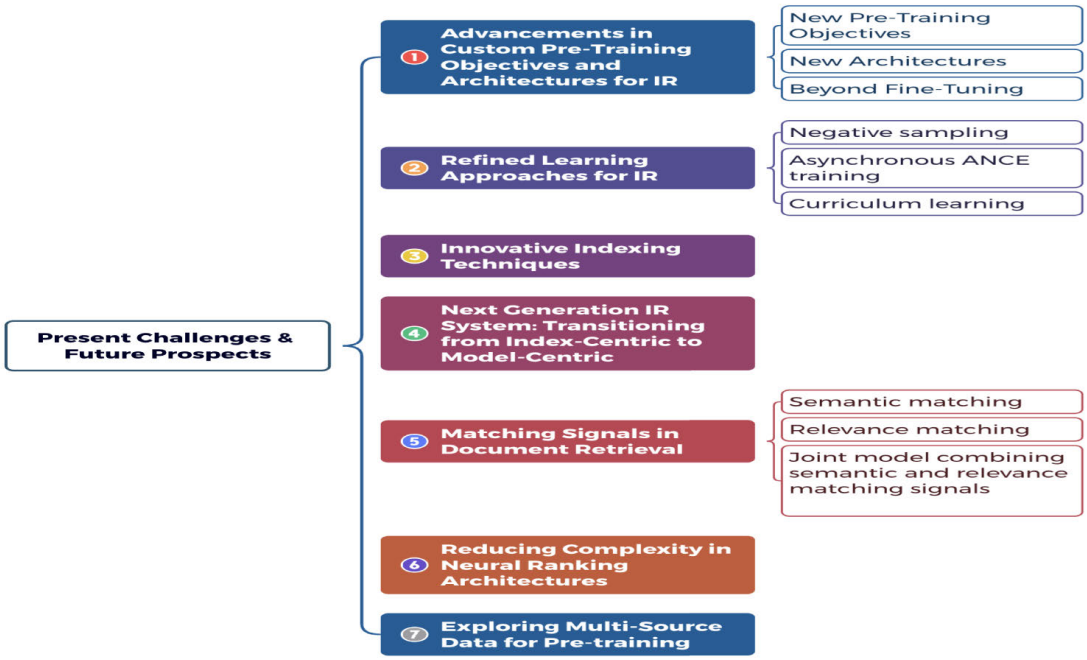

FIGURE 8. Highlights the main theme focused on unresolved issues and potential future developments in the context of IR.

significantly improving performance. Designing additional appropriate objectives for IR is still in its early stages.

New Architectures: Additional investigation direction involves devising novel architectures based on specific downstream tasks. For example, Gao and Callan [242] introduced Condenser, an innovative Transformer designed to discourse structural readiness. This approach generates steady advancement over traditional language models and significantly enhances robust task-specific. However, designing a clever pre-training model architecture suitable has not been extensively explored compared to investigating new pre-training objectives for IR.

Beyond Fine-Tuning: Although fine-tuning is the widely used technique for Pretrained Transformer Models (PTMs) for downstream tasks, it contains some weaknesses. It acts inadequately on tasks without sufficient control data to sustain fine-tuning and is inadequate for fine-tuning parameters on every downstream task. Prompt tuning, which designs discrete [243], [244] or continuous [245], [246] prompts for specific downstream tasks, is a good method to decrease the computational expense of utilizing PT models for downstream tasks. Though prompt tuning has gained outstanding outcomes in domains such as data extraction [247], [248], text classification [249], [250], and fact probing [243], [251], there has existed no extended work on prompt tuning.

In summary, limited progress has been made in developing large pre-training objectives and architectures. It is crucial to consider retrieval requirements, such as maximizing the recall of potentially relevant documents and modeling task-dependent characteristics when designing innovative pre-training objectives for IR.

B. REFINED LEARNING APPROACHES FOR IR

Creating benchmark datasets for information retrieval tasks involves pooling, leading to a known bias issue [91], [98]. Addressing this problem requires intelligent learning techniques, such as debiased contrastive objectives [252]. Hard negative instances could enhance the model's distinguishing capability, but strategies for mining them have not been thoroughly explored. Asynchronous ANCE training [95] is a prominent method but has limitations due to the increased training cost. Research suggests that learning with hard and easy negative samples is more effective [101], and exploring sophisticated training techniques, like curriculum learning [253], is valuable. Supervised data is often limited and prone to long-tail and sparsity problems. Weakly unsupervised or supervised learning methods offer promising directions, such as contrastive learning [253], [254]. Content-weak supervision strategies leverage the inner structure of documents to extract training labels [256].

C. INNOVATIVE INDEXING TECHNIQUES

Indexing schemes are crucial in IR tasks, as they dictate the organization and retrieval of extensive big documents. Dense retrieval techniques that comprehend dense representations for queries and documents typically depend on the ANN technique for better vector search in internet search [109], [257]. Present dense retrieval techniques often divide representation learning and index building, which introduces several disadvantages. The indexing operation is unable to utilize supervised knowledge, and the separately obtained representation and index are not optimally consistent, affecting retrieval performance. Some studies have investigated combined training of indexes and encoders in the recommendation and image retrieval domains [258], [259]. Designing combined learning schemes for retrieval-stage and indexing techniques in information retrieval represents a promising research direction.

Additionally, creating more suitable ANN algorithms to handle big-size documents and sustain better and more accurate retrieval is vital. Generally, two types of ANN algorithms are aimed at enhancing retrieval efficiency: nonspecific ANN search techniques [260], [261] and vector compression techniques [262], [263], [264]. Each method has limitations or drawbacks, such as sizable index sizes for non-exhaustive methods and suboptimal performance for compression methods. With the growing significance of dense retrieval methods, there is an urgent need to develop state-of-the-art ANN algorithms to balance efficiency and effectiveness better.

D. EXPLORING MULTI-SOURCE DATA FOR PRE-TRAINING

Multi-modal creating PTMs based on cross-modal data, such as text, image, audio, and video. While progress has been made in vision-language pre-training (VLP) for various downstream tasks, existing models are not thoroughly evaluated for IR tasks. The future of multi-modal pre-training should explore better vision-language objectives, integrating more modalities and data for IR tasks [265], [266], [267].

Multi-lingual pre-training addresses the need for PTMs that operate with numerous languages rather than just English. Although some existing multi-lingual PTMs show language transfer abilities, they are primarily designed for NLP tasks and not cross-lingual tasks in IR. Future research should focus on models better suited for cross-lingual IR tasks [5], [268], [269].

Knowledge-improved pre-training involves integrating external knowledge, such as knowledge graphs and field-specific data, into Pretrained Transformer Models (PTMs) to enhance IR performance. While there has been significant work in this area, we observe that there is still potential for these techniques to be more specifically and effectively tailored to the unique challenges and requirements of IR tasks. Future research should explore more efficient ways to model this knowledge for IR, as well as strive for greater interpretability in modeling knowledge for downstream tasks [270], [271], [272], [273], [274].

E. NEXT GENERATION IR SYSTEM: SHIFTING FROM INDEX-CENTRIC TO MODEL-CENTRIC

With the remarkable advancements within PTMs, traditional multi-stage systems are index-centric. Still, pre-trained models with large sizes can encode extensive world knowledge. This capability may allow them to produce direct results in response to information needs. Metzler et al. [275] presented a concept to construct model-based systems using powerful pre-trained models. This concept incorporates an index during training. Although this presents an intriguing vision, it remains somewhat abstract. Tay et al. [276] developed a new paradigm established on the T5 model, achieving notable enhancement by training with indexing and retrieval in a multi-task design. Likewise, Zhou et al. [277] introduced Dynamic Retriever, a model-based IR system built on BERT. The BERT-based dense retriever is initially fine-tuned with query-document pairs, and then model parameters are initialized using the generated document embeddings. The model is further fine-tuned with query-docid pairs. Despite these initial explorations, numerous challenges still need to be addressed. For example, how can we create semantics-based document identifications, and how should the model be updated when the document collection changes?

F. ESSENTIAL MATCHING SIGNALS IN DOCUMENT RETRIEVAL

In information retrieval, neural ranking models concentrate on two key matching approaches, relevance, and semantic [21]. Semantic matching is utilized to compare a query and a document. However, relying solely on semantic matching is inadequate for document retrieval, particularly when queries contain specific keywords.

Relevance matching [21] tackles the heterogeneity of queries and documents in ad-hoc document retrieval. Conven-

tional retrieval models, such as BM25, mainly employ exact matching to rank documents that can be incorporated with neural ranking models to boost retrieval effectiveness.

Both relevance and semantic signals are crucial to address diverse scenarios in ad-hoc retrieval tasks. The joint model proposed by [24] merges the [CLS] representation from BERT with existing relevance-based neural ranking models, providing relevance and semantic signals. Nonetheless, the length constraint of BERT presents challenges during training and inference, necessitating the splitting of documents into sentences or passages and increasing training and inference durations.

G. REDUCING COMPLEXITY IN NEURAL RANKING ARCHITECTURES

Employing BERT as a semantic matching component has drawbacks, including the BERT length limit and computational complexity. The maximum token count allowed by BERT is considerably lower than the average document length, which presents a challenge when processing longer texts. Additionally, BERT's computational complexity can increase processing time and resource requirements. Researchers can explore selection techniques for sentences or passages to address this issue. Models like Colbert [257], RepBERT [91], and RocketQA [98] have been proposed to overcome these limitations. Vector compression methods, such as PQ [264] and LSH [263], have been integrated into neural ranking models. Hofstätter et al. [118] and Kitaev et al. [278] introduced methods to reduce the complexity of Transformers, while ElMo can provide deep contextualized embeddings without a length limit constraint.

VII. CONCLUSION

This review offers an extensive analysis of the current state of information retrieval models/techniques. Spanning from early semantic retrieval techniques to the latest advancements in neural semantic retrieval approaches, we delve into the intricate relationships between these methods. Our examination is structured around fundamental IR topics, such as first-stage and second-stage retrieval and learning neural semantic retrieval models. Furthermore, the review emphasizes this domain's primary obstacles and challenges, shedding light on potential future research directions. As a whole, this review aims to serve as a valuable resource for researchers intrigued by this demanding subject, fostering innovative thinking and the progression of the field.

REFERENCES

[1] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, vol. 520. Reading, MA, USA: Addison-Wesley, 2010.
[2] A. Yates, R. Nogueira, and J. Lin, "Pretrained transformers for text ranking: BERT and beyond," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 1154–1156.
[3] C. Yahui, "Convolutional neural network for sentence classification," M.S. thesis, Univ. Waterloo, 2015.
[4] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, 2015, pp. 1–9.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[6] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI*, vol. 1, no. 8, p. 9, 2019.
[8] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.
[9] F. Du, J. Zhang, J. Hu, and R. Fei, "Discriminative multi-modal deep generative models," *Knowl.-Based Syst.*, vol. 173, pp. 74–82, Jun. 2019.
[10] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[13] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," *Inf. Process. Manag.*, vol. 57, no. 6, Nov. 2020, Art. no. 102067.
[14] K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. de Rijke, and M. Lease, "Neural information retrieval: At the end of the early years," *Inf. Retr. J.*, vol. 21, nos. 2–3, pp. 111–182, Jun. 2018.
[15] Y. Fan, X. Xie, Y. Cai, J. Chen, X. Ma, X. Li, R. Zhang, and J. Guo, "Pre-training methods in information retrieval," *Found. Trends Inf. Retr.*, vol. 16, no. 3, pp. 178–317, 2022.
[16] H. Li and J. Xu, "Semantic matching in search," *Found. Trends Inf. Retr.*, vol. 7, no. 5, pp. 343–469, 2014.
[17] J. Lin, R. Nogueira, and A. Yates, "Pretrained transformers for text ranking: BERT and beyond," *Synth. Lectures Hum. Lang. Technol.*, vol. 14, no. 4, pp. 1–325, Oct. 2021.
[18] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. NIST*, vol. 109, 1995, p. 109.
[19] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
[20] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–8.
[21] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2016, pp. 55–64.
[22] B. Mitra, F. Diaz, and N. Craswell, "Learning to match using local and distributed representations of text for web search," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1291–1299.
[23] J. Zhou and E. Agichtein, "RLIRank: Learning to rank with reinforcement learning for dynamic search," in *Proc. Web Conf.*, Apr. 2020, pp. 2842–2848.
[24] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "CEDR: Contextualized embeddings for document ranking," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 1101–1104.
[25] J. Li, H. Zeng, L. Peng, J. Zhu, and Z. Liu, "Learning to rank method combining multi-head self-attention with conditional generative adversarial nets," *Array*, vol. 15, Sep. 2022, Art. no. 100205.
[26] A. Boubacar and Z. Niu, "Concept based query expansion," in *Proc. 9th Int. Conf. Semantics, Knowl. Grids*, Oct. 2013, pp. 198–201.
[27] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proc. 17th Annu. Int. ACM-SIGIR Conf. Res. Develop. Inf. Retr.* London, U.K.: Springer, 1994, pp. 61–69.
[28] J. Bai, J.-Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 15–22.
[29] J. John, "Relevance feedback in information retrieval," in *The SMART Retrieval System: Experiments in Automatic Document Processing*. 1971.
[30] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. 10th Int. Conf. Inf. Knowl. Manag.*, Oct. 2001, pp. 403–410.

[31] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2008, pp. 243–250.

[32] Y. Lv and C. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proc. 18th ACM Conf. Inf. Knowl. Manag.*, Nov. 2009, pp. 1895–1898.

[33] H. Zamani, J. Dadashkarimi, A. Shakery, and W. B. Croft, "Pseudo-relevance feedback based on matrix factorization," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2016, pp. 1483–1492.

[34] O. Kurland and L. Lee, "Corpus structure, language models, and ad hoc information retrieval," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2004, pp. 194–201.

[35] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2004, pp. 186–193.

[36] B. Billerbeck and J. Zobel, "Document expansion versus query expansion for ad-hoc retrieval," in *Proc. 10th Australas. Document Comput. Symp.*, 2005, pp. 34–41.

[37] T. Tao, X. Wang, Q. Mei, and C. Zhai, "Language model information retrieval with document expansion," in *Proc. Conf. Human Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2006, pp. 407–414.

[38] E. Agirre, X. Arregi, and A. Otegi, "Document expansion based on WordNet for robust IR," in *Proc. Coling*, 2010, pp. 9–17.

[39] M. Efron, P. Organisciak, and K. Fenlon, "Improving retrieval of short texts through document expansion," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2012, pp. 911–920.

[40] G. Sherman and M. Efron, "Document expansion using external collections," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1045–1048.

[41] J. L. Fagan, *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Nonsyntactic Methods*. Cornell Univ., 1988.

[42] M. Mitra, C. Buckley, A. Singhal, and C. Cardie, "An analysis of statistical and syntactic phrases," in *Proc. RIAO*, vol. 97, 1997, pp. 200–214.

[43] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proc. 8th Int. Conf. Inf. Knowl. Manag.*, 1999, pp. 316–321.

[44] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments: Part 2," *Inf. Process. Manag.*, vol. 36, no. 6, pp. 809–840, Nov. 2000.

[45] R. Nallapati and J. Allan, "Capturing term dependencies using a language model based on sentence trees," in *Proc. 11th Int. Conf. Inf. Knowl. Manag.*, Nov. 2002, pp. 383–390.

[46] J. Gao, J.-Y. Nie, G. Wu, and G. Cao, "Dependence language model for information retrieval," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2004, pp. 170–177.

[47] J. Xu, H. Li, and C. Zhong, "Relevance ranking using kernels," in *Proc. 6th Asia Inf. Retr. Societies Conf.* Berlin, Germany: Springer, 2010, pp. 1–12.

[48] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, "Generalized vector spaces model in information retrieval," in *Proc. 8th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. - SIGIR*, 1985, pp. 18–25.

[49] F. Diaz, "Regularizing ad hoc retrieval scores," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2005, pp. 672–679.

[50] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2006, pp. 178–185.

[51] X. Yi and A. James, "A comparative study of utilizing topic models for information retrieval," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, 2009, pp. 29–41.

[52] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Inf. Retr.*, vol. 14, no. 2, pp. 178–203, Apr. 2011.

[53] A. Atreya and C. Elkan, "Latent semantic indexing (LSI) fails for TREC collections," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 5–10, Mar. 2011.

[54] A. Berger and J. Lafferty, "Information retrieval as statistical translation," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 219–226, Aug. 2017.

[55] M. Karimzadehgan and C. Zhai, "Estimation of statistical translation models based on mutual information for ad hoc information retrieval," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2010, pp. 323–330.

[56] J. Gao, X. He, and J.-Y. Nie, "Clickthrough-based translation models for web search: From word models to phrase models," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2010, pp. 1139–1148.

[57] M. Karimzadehgan and C. Zhai, "Axiomatic analysis of translation language model for information retrieval," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, 2012, pp. 268–280.

[58] S. Riezler and Y. Liu, "Query rewriting using monolingual statistical machine translation," *Comput. Linguistics*, vol. 36, no. 3, pp. 569–582, Sep. 2010.

[59] J. Gao and J.-Y. Nie, "Towards concept-based translation models using search logs for query expansion," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2012, pp. 1–10.

[60] G. Zheng and J. Callan, "Learning to reweight terms with distributed representations," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 575–584.

[61] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi, "Integrating and evaluating neural word embeddings in information retrieval," in *Proc. 20th Australas. Document Comput. Symp.*, Dec. 2015, pp. 1–8.

[62] Z. Dai and J. Callan, "Context-aware sentence/passage term importance estimation for first stage retrieval," 2019, *arXiv:1910.10687*.

[63] J. Frej, P. Mulhem, D. Schwab, and J.-P. Chevallet, "Learning term discrimination," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1993–1996.

[64] Z. Dai and J. Callan, "Context-aware term weighting for first stage passage retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1533–1536.

[65] J. Mackenzie, Z. Dai, L. Gallagher, and J. Callan, "Efficiency implications of term weighting for passage retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1821–1824.

[66] J. Lin and X. Ma, "A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques," 2021, *arXiv:2106.14807*.

[67] R. Nogueira, W. Yang, J. Lin, and K. Cho, "Document expansion by query prediction," 2019, *arXiv:1904.08375*.

[68] R. Nogueira, J. Lin, and A. I. Epistemic, "From doc2query to docTTTTTquery," Tech. Rep., 2019.

[69] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, "Generation-augmented retrieval for open-domain question answering," 2020, *arXiv:2009.08553*.

[70] M. Yan, C. Li, B. Bi, W. Wang, and S. Huang, "A unified pretraining framework for passage ranking and expansion," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 5, pp. 4555–4563.

[71] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Expansion via prediction of importance with contextualization," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1573–1576.

[72] Y. Bai, X. Li, G. Wang, C. Zhang, L. Shang, J. Xu, Z. Wang, F. Wang, and Q. Liu, "SparTerm: Learning term-based sparse representation for fast text retrieval," 2020, *arXiv:2010.00768*.

[73] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, "SPLADE v2: Sparse lexical and expansion model for information retrieval," 2021, *arXiv:2109.10086*.

[74] A. Mallia, O. Khattab, T. Suel, and N. Tonellotto, "Learning passage impacts for inverted indexes," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1723–1727.

[75] S. Zhuang and G. Zuccon, "Fast passage re-ranking with contextualized exact term matching and efficient passage expansion," 2021, *arXiv:2108.08513*.

[76] E. Choi, S. Lee, M. Choi, H. Ko, Y.-I. Song, and J. Lee, "SpaDE: Improving sparse representations using a dual document encoder for first-stage retrieval," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2022, pp. 272–282.

[77] R. Salakhutdinov and G. Hinton, "Semantic hashing," *Int. J. Approx. Reasoning*, vol. 50, no. 7, pp. 969–978, Jul. 2009.

[78] H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, and J. Kamps, "From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2018, pp. 497–506.

[79] K.-R. Jang, J. Kang, G. Hong, S.-H. Myaeng, J. Park, T. Yoon, and H. Seo, "Ultra-high dimensional sparse representations with binarization for efficient text retrieval," 2021, *arXiv:2104.07198*.

[80] I. Yamada, A. Asai, and H. Hajishirzi, "Efficient passage retrieval with hashing for open-domain question answering," 2021, *arXiv:2106.00882*.

[81] C. Lassance, T. Formal, and S. Clinchant, "Composite code sparse autoencoders for first stage retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2136–2140.
[82] S. Clinchant and P. Florent, "Aggregating continuous word embeddings for information retrieval," in *Proc. Workshop Continuous Vector Space Models Their Compositionality*, 2013, pp. 100–109.
[83] I. Vulic and M.-F. Moens, "Monolingual and cross-lingual information retrieval models based on (Bilingual) word embeddings," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 363–372.
[84] T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2015, pp. 1411–1420.
[85] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana, "A dual embedding space model for document ranking," 2016, *arXiv:1602.01137*.
[86] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, "Efficient natural language response suggestion for smart reply," 2017, *arXiv:1705.00652*.
[87] D. Gillick, A. Presta, and G. S. Tomar, "End-to-end retrieval in continuous space," 2018, *arXiv:1811.08008*.
[88] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," 2020, *arXiv:2004.04906*.
[89] M. Seo, T. Kwiatkowski, A. P. Parikh, A. Farhadi, and H. Hajishirzi, "Phrase-indexed question answering: A new challenge for scalable document comprehension," 2018, *arXiv:1804.07726*.
[90] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. T. Yih, T. Rocktaschel, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.
[91] J. Zhan, J. Mao, Y. Liu, M. Zhang, and S. Ma, "RepBERT: Contextualized text embeddings for first-stage retrieval," 2020, *arXiv:2006.15498*.
[92] M. Wrzalik and D. Krechel, "CoRT: Complementary rankings from transformers," 2020, *arXiv:2010.10252*.
[93] P. Nie, Y. Zhang, X. Geng, A. Ramamurthy, L. Song, and D. Jiang, "DC-BERT: Decoupling question and document for efficient contextual encoding," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1829–1832.
[94] Y. Yang, N. Jin, K. Lin, M. Guo, and D. Cer, "Neural retrieval for question answering with cross-attention supervised data augmentation," 2020, *arXiv:2009.13815*.
[95] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk, "Approximate nearest neighbor negative contrastive learning for dense text retrieval," 2020, *arXiv:2007.00808*.
[96] J. Zhan, J. Mao, Y. Liu, M. Zhang, and S. Ma, "Learning to retrieve: How to train a dense retrieval model effectively and efficiently," 2020, *arXiv:2010.10469*.
[97] X. Shan, C. Liu, Y. Xia, Q. Chen, Y. Zhang, K. Ding, Y. Liang, A. Luo, and Y. Luo, "GLOW : Global weighted self-attention network for web search," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021, pp. 519–528.
[98] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, "RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering," 2020, *arXiv:2010.08191*.
[99] J. Lee, M. Sung, J. Kang, and D. Chen, "Learning dense representations of phrases at scale," 2020, *arXiv:2012.12624*.
[100] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury, "Efficiently teaching an effective dense retriever with balanced topic aware sampling," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 113–122.
[101] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma, "Optimizing dense retrieval model training with hard negatives," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1503–1512.
[102] S. Yu, Z. Liu, C. Xiong, T. Feng, and Z. Liu, "Few-shot conversational dense retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 829–838.
[103] Y. Li, Z. Liu, C. Xiong, and Z. Liu, "More robust dense retrieval with contrastive dual learning," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Jul. 2021, pp. 287–296.
[104] R. Ren, S. Lv, Y. Qu, J. Liu, W. X. Zhao, Q. She, H. Wu, H. Wang, and J.-R. Wen, "PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval," 2021, *arXiv:2108.06027*.
[105] O. Khattab, C. Potts, and M. Zaharia, "Relevance-guided supervision for OpenQA with ColBERT," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 929–944, Sep. 2021.
[106] D. Singh, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, "End-to-end training of multi-document reader and retriever for open-domain question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25968–25981.
[107] H. Yu, C. Xiong, and J. Callan, "Improving query representations for dense retrieval with pseudo relevance feedback," 2021, *arXiv:2108.13454*.
[108] X. Wang, C. Macdonald, N. Tonellotto, and I. Ounis, "Pseudo-relevance feedback for multiple representation dense retrieval," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Jul. 2021, pp. 297–306.
[109] Y. Cai, Y. Fan, J. Guo, R. Zhang, Y. Lan, and X. Cheng, "A discriminative semantic ranker for question retrieval," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Jul. 2021, pp. 251–260.
[110] B. Wu, Z. Zhang, J. Wang, and H. Zhao, "Sentence-aware contrastive learning for open-domain passage retrieval," 2021, *arXiv:2110.07524*.
[111] R. Ren, Y. Qu, J. Liu, W. Xin Zhao, Q. She, H. Wu, H. Wang, and J.-R. Wen, "RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking," 2021, *arXiv:2110.07367*.
[112] E. Lindgren, S. Reddi, R. Guo, and S. Kumar, "Efficient training of retrieval models using negative cache," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 4134–4146.
[113] J. Lu, G. Hernandez Abrego, J. Ma, J. Ni, and Y. Yang, "Multi-stage training with improved negative contrast for neural passage retrieval," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6091–6103.
[114] S.-C. Lin, J.-H. Yang, and J. Lin, "Distilling dense representations for ranking using tightly-coupled teachers," 2020, *arXiv:2010.11386*.
[115] A. V. Tahami, K. Ghajar, and A. Shakery, "Distilling knowledge for fast retrieval-based chat-bots," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2081–2084.
[116] G. Izacard and E. Grave, "Distilling knowledge from reader to retriever for question answering," 2020, *arXiv:2012.04584*.
[117] W. Lu, J. Jiao, and R. Zhang, "TwinBERT: Distilling knowledge to twin-structured compressed BERT models for large-scale retrieval," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 2645–2652.
[118] S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, and A. Hanbury, "Improving efficient neural ranking models with cross-architecture knowledge distillation," 2020, *arXiv:2010.02666*.
[119] S. Yang and M. Seo, "Is retriever merely an approximator of reader?" 2020, *arXiv:2010.10999*.
[120] J. Choi, E. Jung, J. Suh, and W. Rhee, "Improving bi-encoder document ranking models with two rankers and multi-teacher distillation," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2192–2196.
[121] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, "Word embedding based generalized language model for information retrieval," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 795–798.
[122] D. Roy, D. Ganguly, M. Mitra, and G. J. F. Jones, "Representing documents and queries as sets of word embedded vectors for information retrieval," 2016, *arXiv:1606.07869*.
[123] L. Boytsov, D. Novak, Y. Malkov, and E. Nyberg, "Off the beaten path: Let's replace term-based retrieval with k-NN search," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2016, pp. 1099–1108.
[124] C. Dos Santos, L. Barbosa, D. Bogdanova, and B. Zadrozny, "Learning hybrid representations to retrieve semantically equivalent questions," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 694–699.
[125] M. Seo, J. Lee, T. Kwiatkowski, A. P. Parikh, A. Farhadi, and H. Hajishirzi, "Real-time open-domain question answering with dense-sparse phrase index," 2019, *arXiv:1906.05807*.
[126] J. Lee, M. Seo, H. Hajishirzi, and J. Kang, "Contextualized sparse representations for real-time open-domain question answering," 2019, *arXiv:1911.02896*.
[127] L. Gao, Z. Dai, T. Chen, Z. Fan, B. Van Durme, and J. Callan, "Complement lexical retrieval model with semantic residual embeddings," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2021, pp. 146–160.
[128] S. Kuzi, M. Zhang, C. Li, M. Bendersky, and M. Najork, "Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach," 2020, *arXiv:2010.01195*.
[129] X. Chen, B. He, K. Hui, Y. Wang, L. Sun, and Y. Sun, "Contextualized offline relevance weighting for efficient and effective neural retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1617–1621.

[130] N. Arabzadeh, X. Yan, and C. L. A. Clarke, "Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2021, pp. 2862–2866.

[131] J. Leonhardt, K. Rudra, M. Khosla, A. Anand, and A. Anand, "Efficient neural ranking using forward indexes," 2021, *arXiv:2110.06051*.

[132] S.-C. Lin and J. Lin, "Densifying sparse representations for passage retrieval by representational slicing," 2021, *arXiv:2112.04666*.

[133] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1999, pp. 34–41.

[134] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[135] C. D. Manning, *An Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[136] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1–7.

[137] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2011, pp. 685–694.

[138] A. Berger and J. Lafferty, "Information retrieval as statistical translation," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1999, pp. 219–226.

[139] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 595–603.

[140] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[141] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[142] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[143] H. Nam, J. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2156–2164.

[144] B. Kong, X. Wang, J. Bai, Y. Lu, F. Gao, K. Cao, J. Xia, Q. Song, and Y. Yin, "Learning tree-structured representation for 3D coronary artery segmentation," *Computerized Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101688.

[145] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[146] W. Zhang, X. Zhao, L. Zhao, D. Yin, G. H. Yang, and A. Beutel, "Deep reinforcement learning for information retrieval: Fundamentals and advances," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2468–2471.

[147] B. Kratzwald, A. Eigenmann, and S. Feuerriegel, "RankQA: Neural question answering with answer re-ranking," 2019, *arXiv:1906.03008*.

[148] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 557–570, Dec. 2018.

[149] D. H. Park and R. Chiba, "A neural language model for query auto-completion," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1189–1192.

[150] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li, "DRN: A deep reinforcement learning framework for news recommendation," in *Proc. World Wide Web Conf.*, 2018, pp. 167–176.

[151] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[152] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 739–744.

[153] F. Scarselli, S. L. Yong, M. Gori, M. Hagenbuchner, A. C. Tsoi, and M. Maggini, "Graph neural networks for ranking web pages," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, 2005, pp. 666–672.

[154] G. M. Correia and A. F. T. Martins, "A simple and effective approach to automatic post-editing with transfer learning," 2019, *arXiv:1906.06253*.

[155] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[156] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[157] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 793–803.

[158] S. Ge, C. Wu, F. Wu, T. Qi, and Y. Huang, "Graph enhanced representation learning for news recommendation," in *Proc. Web Conf.*, Apr. 2020, pp. 2863–2869.

[159] S. Kim, A. S. Rawat, M. Zaheer, S. Jayasumana, V. Sadhanala, W. Jitkrittum, A. K. Menon, R. Fergus, and S. Kumar, "EmbedDistill: A geometric knowledge distillation for information retrieval," 2023, *arXiv:2301.12005*.

[160] N. Fuhr, "Optimum polynomial retrieval functions based on the probability ranking principle," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 183–204, Jul. 1989.

[161] S. K. M. Wong, Y. J. Cai, and Y. Y. Yao, "Computation of term associations by a neural network," in *Proc. 16th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1993, pp. 107–115.

[162] F. C. Gey, "Inferring probability of relevance using the method of logistic regression," in *Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1994, pp. 222–231.

[163] H. Schutze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*. 2008.

[164] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[165] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2002, pp. 133–142.

[166] F. Radlinski and T. Joachims, "Query chains: Learning to rank from implicit feedback," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 239–248.

[167] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2007.

[168] H. Li, *Learning to Rank for Information Retrieval and Natural Language Processing*, 2022.

[169] C. J. Burges, "From RankNet to LambdaRank to LambdaMART: An overview," *Learning*, vol. 11, p. 81, Jun. 2010.

[170] Y. Ganjisaffar, R. Caruana, and C. V. Lopes, "Bagging gradient-boosted trees for high precision, low variance ranking models," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2011, pp. 85–94.

[171] R. K. Pasumarthi, S. Bruch, X. Wang, C. Li, M. Bendersky, M. Najork, J. Pfeifer, N. Golbandi, R. Anil, and S. Wolf, "TF-ranking: Scalable TensorFlow library for learning-to-rank," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2970–2978.

[172] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees, "Overview of the TREC 2019 deep learning track," 2020, *arXiv:2003.07820*.

[173] B. Mitra and N. Craswell, "An introduction to neural information retrieval T," *Found. Trends Inf. Retr.*, vol. 13, no. 1, pp. 1–126, 2018.

[174] J. Xu, X. He, and H. Li, "Deep learning for matching in search and recommendation," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 1365–1368.

[175] J. Bromley, I. Guyon, Y. LeCun, E. Sackinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1993, pp. 1–8.

[176] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag.*, 2013, pp. 2333–2338.

[177] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 373–374.

[178] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[179] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1–7.

[180] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag.*, Nov. 2014, pp. 101–110.

[181] J. Mueller and T. Aditya, "Siamese recurrent architectures for learning sentence similarity," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–7.

[182] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[183] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[184] S. Wan, L. Yanyan, G. Jiafeng, X. Jun, P. Liang, and C. Xueqi, "A deep architecture for semantic matching with multiple positional sentence representations," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–7.

[185] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–10.

[186] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*.

[187] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana, "Improving document ranking with dual word embeddings," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 83–84.

[188] Q. Le and M. Tomas, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[189] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 55–64.

[190] Y. Fan, J. Guo, Y. Lan, J. Xu, C. Zhai, and X. Cheng, "Modeling diverse relevance patterns in ad-hoc retrieval," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 375–384.

[191] L. Yang, Q. Ai, J. Guo, and W. B. Croft, "ANMM: Ranking short answer texts with attention-based neural matching model," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2016, pp. 287–296.

[192] L. Pang, Y. Lan, J. Guo, J. Xu, and X. Cheng, "A study of MatchPyramid models on ad-hoc retrieval," 2016, *arXiv:1606.04648*.

[193] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, and X. Cheng, "DeepRank: A new deep architecture for relevance ranking in information retrieval," in *Proc. ACM Conf. Inf. Knowl. Manag.*, Nov. 2017, pp. 257–266.

[194] H. He and J. Lin, "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 937–948.

[195] A. Jaech, H. Kamisetty, E. Ringger, and C. Clarke, "Match-tensor: A deep relevance model for search," 2017, *arXiv:1701.07795*.

[196] R. Abri, S. Abri, and S. Cetin, "Providing a topic-based LSTM model to re-rank search results," in *Proc. 7th Int. Conf. Mach. Learn. Technol. (ICMLT)*, Mar. 2022, pp. 249–254.

[197] S. Wan, Y. Lan, J. Xu, J. Guo, L. Pang, and X. Cheng, "Match-SRNN: Modeling the recursive matching structure with spatial RNN," 2016, *arXiv:1604.04378*.

[198] K. Hui, A. Yates, K. Berberich, and G. de Melo, "PACRR: A position-aware neural IR model for relevance matching," 2017, *arXiv:1704.03940*.

[199] R. McDonald, G.-I. Brokos, and I. Androutsopoulos, "Deep relevance ranking using enhanced document-query interactions," 2018, *arXiv:1809.01682*.

[200] Z. Dai, C. Xiong, J. Callan, and Z. Liu, "Convolutional neural networks for soft-matching N-grams in ad-hoc search," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 126–134.

[201] W. Lan and W. Xu, "Character-based neural networks for sentence pair modeling," 2018, *arXiv:1805.08297*.

[202] Y. Nie, Y. Li, and J.-Y. Nie, "Empirical study of multi-level convolution models for IR based on representations and interactions," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Sep. 2018, pp. 59–66.

[203] Z. Tang and H. Y. Grace, "DeepTileBars: Visualizing term distribution for neural information retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 289–296.

[204] B. Mitra and N. Craswell, "An updated duet model for passage re-ranking," 2019, *arXiv:1903.07666*.

[205] J. Lin, "The neural hype and comparisons against weak baselines," in *Proc. ACM SIGIR Forum*, New York, NY, USA, 2019, pp. 40–51.

[206] W. Yang, K. Lu, P. Yang, and J. Lin, "Critically examining the 'neural hype' weak baselines and the additivity of effectiveness gains from neural ranking models," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2019, pp. 1129–1132.

[207] S. Kim, K. Inho, and K. Nojun, "Semantic sentence matching with densely-connected recurrent and co-attentive information," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6586–6593.

[208] K. Zhang, G. Lv, L. Wang, L. Wu, E. Chen, F. Wu, and X. Xie, "DRr-Net: Dynamic re-read network for sentence semantic matching," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 7442–7449.

[209] C. Tan, F. Wei, W. Wang, W. Lv, and M. Zhou, "Multiway attention networks for modeling sentence pairs," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4411–4417.

[210] S. Wang and J. Jiang, "A compare-aggregate model for matching text sequences," 2016, *arXiv:1611.01747*.

[211] W. Yin, H. Schutze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 259–272, Dec. 2016.

[212] W. Yang, H. Zhang, and J. Lin, "Simple applications of BERT for ad hoc document retrieval," 2019, *arXiv:1903.10972*.

[213] H. Zhang, M. Abualsaud, N. Ghelani, M. D. Smucker, G. V. Cormack, and M. R. Grossman, "Effective user interaction for high-recall retrieval: Less is more," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2018, pp. 187–196.

[214] Z. Yang, Z. Dai, Y. Yang, Y. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[215] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.

[216] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of BERT in ranking," 2019, *arXiv:1904.07531*.

[217] Z. Dai and J. Callan, "Deeper text understanding for IR with contextual neural language modeling," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 985–988.

[218] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-stage document ranking with BERT," 2019, *arXiv:1910.14424*.

[219] L. Boualili, J. G. Moreno, and M. Boughanem, "MarkedBERT: Integrating traditional IR cues in pre-trained language models for passage retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1977–1980.

[220] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," Tech. Rep., 2016.

[221] R. Nogueira and K. Cho, "Passage re-ranking with BERT," 2019, *arXiv:1901.04085*.

[222] C. Li, Y. Sun, B. He, L. Wang, K. Hui, A. Yates, L. Sun, and J. Xu, "NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval," 2018, *arXiv:1810.12936*.

[223] L. Lu, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, and P. Mooney, "CORD-19: The COVID-19 open research dataset," 2020, *arXiv:2004.10706*.

[224] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, and Y. Almirantis, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–28, Dec. 2015.

[225] V. Boteva, D. Gholipour, A. Sokolov, and S. Riezler, "A full-text learning to rank dataset for medical information retrieval," in *Proc. Eur. Conf. Inf. Retr.* Padua, Italy: Springer, 2016, pp. 716–722.

[226] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 453–466, Nov. 2019.

[227] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," 2018, *arXiv:1809.09600*.

[228] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur, "WWW'18 open challenge: Financial opinion mining and question answering," in *Proc. Companion Web Conf.*, 2018, pp. 1941–1942.

[229] T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal, "QASC: A dataset for question answering via sentence composition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8082–8090.

[230] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: A study and an open task," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 813–820.

[231] H. Wachsmuth, S. Syed, and B. Stein, "Retrieval of the best counterargument without prior topic knowledge," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 241–251.

[232] A. Bondarenko, L. Gienapp, M. Frobe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, and M. Hagen, "Overview of touche 2021: Argument retrieval," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Berlin, Germany: Springer, 2021, pp. 450–467.

[233] F. Hasibi, F. Nikolaev, C. Xiong, K. Balog, S. E. Bratsberg, A. Kotov, and J. Callan, "DBpedia-entity v2: A test collection for entity search," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1265–1268.

[234] T. Qin and T.-Y. Liu, "Introducing LETOR 4.0 datasets," 2013, *arXiv:1306.2597*.

[235] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Linguistic knowledge enhanced language representation for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6975–6988.

[236] A. Sahu and S. G. Sanjeevi, "Better fine-tuning with extracted important sentences for abstractive summarization," in *Proc. Int. Conf. Commun., Control Inf. Sci. (ICCISc)*, Jun. 2021, pp. 11328–11339.

[237] K. Lee, M.-W. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," 2019, *arXiv:1906.00300*.

[238] W.-C. Chang, F. X. Yu, Y.-W. Chang, Y. Yang, and S. Kumar, "Pre-training tasks for embedding-based large-scale retrieval," 2020, *arXiv:2002.03932*.

[239] K. Guu, K. Lee, K. Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[240] X. Ma, J. Guo, R. Zhang, Y. Fan, X. Ji, and X. Cheng, "PROP: Pre-training with representative words prediction for ad-hoc retrieval," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 283–291.

[241] X. Ma, J. Guo, R. Zhang, Y. Fan, Y. Li, and X. Cheng, "B-PROP: Bootstrapped pre-training with representative words prediction for ad-hoc retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1513–1522.

[242] L. Gao and J. Callan, "Condenser: A pre-training architecture for dense retrieval," 2021, *arXiv:2104.08253*.

[243] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" 2019, *arXiv:1909.01066*.

[244] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020, *arXiv:2012.15723*.

[245] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," 2021, *arXiv:2103.10385*.

[246] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021, *arXiv:2104.08691*.

[247] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen, "KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2778–2788.

[248] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt tuning with rules for text classification," *AI Open*, vol. 3, pp. 182–192, 2022.

[249] R. Puri and B. Catanzaro, "Zero-shot text classification with generative language models," 2019, *arXiv:1912.10165*.

[250] T. Schick and H. Schutze, "Exploiting cloze questions for few shot text classification and natural language inference," 2020, *arXiv:2001.07676*.

[251] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Trans. Assoc. for Comput. Linguistics*, vol. 8, pp. 423–438, Dec. 2020.

[252] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8765–8775.

[253] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[254] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[255] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[256] Z. Dai and J. Callan, "Context-aware document term weighting for ad-hoc search," in *Proc. Web Conf.*, Apr. 2020, pp. 1897–1907.

[257] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2020, pp. 39–48.

[258] T. Yu, J. Yuan, C. Fang, and H. Jin, "Product quantization network for fast image retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 186–201.

[259] H. Zhang, H. Shen, Y. Qiu, Y. Jiang, S. Wang, S. Xu, Y. Xiao, B. Long, and W.-Y. Yang, "Joint learning of deep retrieval model and product quantization based embedding index," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1718–1722.

[260] E. Bernhardsson, "ANNOY: Approximate nearest neighbors in CE++/Python," Tech. Rep., 2018.

[261] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020.

[262] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 744–755, Apr. 2014.

[263] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.

[264] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[265] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 201–216.

[266] Y. Huo, M. Zhang, G. Liu, H. Lu, Y. Gao, G. Yang, J. Wen, H. Zhang, B. Xu, W. Zheng, and Z. Xi, "WenLan: Bridging vision and language by large-scale multi-modal pre-training," 2021, *arXiv:2103.06561*.

[267] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y. C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 565–580.

[268] G. Lample and A. Conneau, "Cross-lingual language model pretraining," 2019, *arXiv:1901.07291*.

[269] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, and M. Zhou, "Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks," 2019, *arXiv:1909.00964*.

[270] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," 2019, *arXiv:1905.07129*.

[271] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "ERNIE: Enhanced representation through knowledge integration," 2019, *arXiv:1904.09223*.

[272] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "KEPLER: A unified model for knowledge embedding and pre-trained language representation," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 176–194, Mar. 2021.

[273] I. Beltagy, K. Lo, and A. Cohan, "A pretrained language model for scientific text," Tech. Rep., 1903.

[274] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[275] D. Metzler, Y. Tay, D. Bahri, and M. Najork, "Rethinking search: Making domain experts out of dilettantes," in *Proc. ACM SIGIR Forum*, vol. 55, 2021, pp. 1–27.

[276] Y. Tay, V. Tran, M. Dehghani, J. Ni, D. Bahri, H. Mehta, Z. Qin, K. Hui, Z. Zhao, J. Gupta, and T. Schuster, "Transformer memory as a differentiable search index," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 21831–21843.
[277] Y. Zhou, J. Yao, Z. Dou, L. Wu, and J.-R. Wen, "DynamicRetriever: A pre-training model-based IR system with neither sparse nor dense index," 2022, *arXiv:2203.00537*.
[278] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020, *arXiv:2001.04451*.
[279] B. Mitra, C. Rosset, D. Hawking, N. Craswell, F. Diaz, and E. Yilmaz, "Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks," 2019, *arXiv:1907.03693*.
[280] B. Mitra, S. Hofstatter, H. Zamani, and N. Craswell, "Conformer-kernel with query term independence for document retrieval," 2020, *arXiv:2007.10434*.
[281] X. Ma, J. Guo, R. Zhang, Y. Fan, and X. Cheng, "Scattered or connected? An optimized parameter-efficient tuning approach for information retrieval," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2022, pp. 1471–1480.
[282] Ö. Sahin, I. Çiçekli, and G. Ercan, "Learning term weights by overfitting pairwise ranking loss," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 5, pp. 1914–1930, Jan. 2022.
[283] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, Aug. 2000.
[284] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus, "ST-MoE: Designing stable and transferable sparse expert models," 2022, *arXiv:2202.08906*.
[285] L. Gao, Z. Dai, and J. Callan, "COIL: Revisit exact lexical match in information retrieval with contextualized inverted list," 2021, *arXiv:2104.07186*.
[286] Q. Cao, H. Trivedi, A. Balasubramanian, and N. Balasubramanian, "DeFormer: Decomposing pre-trained transformers for faster question answering," 2020, *arXiv:2005.00697*.
[287] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Efficient document re-ranking for transformers by precomputing term representations," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 49–58.
[288] Y. Feldman and R. El-Yaniv, "Multi-hop paragraph retrieval for open-domain question answering," 2019, *arXiv:1906.06606*.
[289] Q. Ai, L. Yang, J. Guo, and W. B. Croft, "Analysis of the paragraph vector model for information retrieval," in *Proc. ACM Int. Conf. Theory Inf. Retr.*, Sep. 2016, pp. 133–142.
[290] C. V. Gysel, M. de Rijke, and E. Kanoulas, "Neural vector spaces for unsupervised information retrieval," *ACM Trans. Inf. Syst.*, vol. 36, no. 4, pp. 1–25, Oct. 2018.
[291] M. Agosti, S. Marchesin, and G. Silvello, "Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval," *ACM Trans. Inf. Syst.*, vol. 38, no. 4, pp. 1–48, Oct. 2020.
[292] X. Liu, N. Jian-Yun, and S. Alessandro, "Constraining word embeddings by prior knowledge-application to medical information retrieval," in *Proc. 12th Asia Inf. Retr. Societies Conf.* Beijing, China: Springer, 2016, pp. 155–167.
[293] L. Tamine, L. Soulier, G.-H. Nguyen, and N. Souf, "Offline versus online representation learning of documents using external knowledge," *ACM Trans. Inf. Syst.*, vol. 37, no. 4, pp. 1–34, Oct. 2019.
[294] M. Tan, C. Dos Santos, B. Xiang, and B. Zhou, "LSTM-based deep learning models for non-factoid answer selection," 2015, *arXiv:1511.04108*.
[295] D. Liang, P. Xu, S. Shakeri, C. Nogueira dos Santos, R. Nallapati, Z. Huang, and B. Xiang, "Embedding-based zero-shot retrieval through query generation," 2020, *arXiv:2009.10270*.
[296] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring," 2019, *arXiv:1905.01969*.
[297] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 329–345, Apr. 2021.
[298] H. Tang, X. Sun, B. Jin, J. Wang, F. Zhang, and W. Wu, "Improving document representations by generating pseudo query embeddings for dense retrieval," 2021, *arXiv:2105.03599*.
[299] Y. Zhu, L. Pang, Y. Lan, H. Shen, and X. Cheng, "LoL: A comparative regularization loss over query reformulation losses for pseudo-relevance feedback," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 825–836.
[300] H. Oosterhuis, "Learning-to-rank at the speed of sampling: Plackett–Luce gradient estimation with minimal computational complexity," 2022, *arXiv:2204.10872*.
[301] Y. Jia and H. Wang, "Learning neural ranking models online from implicit user feedback," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 431–441.
[302] X. Wu, Q. Liu, J. Qin, and Y. Yu, "PeerRank: Robust learning to rank with peer loss over noisy labels," *IEEE Access*, vol. 10, pp. 6830–6841, 2022.
[303] Z. Zhang, Y. Zhang, X. Li, Y. Qian, and T. Zhang, "BMCSA: Multi-feature spatial convolution semantic matching model based on BERT," *J. Intell. Fuzzy Syst.*, vol. 43, no. 4, pp. 4083–4093, Aug. 2022.
[304] K. Zhang, G. Lv, L. Wu, E. Chen, Q. Liu, and M. Wang, "LadRa-Net: Locally aware dynamic reread attention net for sentence semantic matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 853–866, Feb. 2023.
[305] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2022.
[306] Accessed: Jul. 18, 2023. [Online]. Available: https://www.microsoft.com/en-us/download/details.aspx?id=52419
[307] [Online]. Available: https://paperswithcode.com/task/information-retrieval

KAILASH A. HAMBARDE received the Ph.D. degree in computer science from S.R.T.M. University, in 2020. He is a Researcher with Universidade da Beira Interior, Portugal.

HUGO PROENÇA (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees, in 2001, 2004, and 2007, respectively. He is currently a Full Professor with the Department of Computer Science, University of Beira Interior, Portugal, has been researching mainly biometrics and visual-surveillance. He is a member of the Editorial Board of the *Image and Vision Computing*, IEEE Access, and *International Journal of Biometrics*. Also, he served as a Guest Editor for the Special Issue of the *Pattern Recognition Letters*, *Image and Vision Computing*, *Signal*, and *Image and Video Processing* journals. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and the Area Editor (ocular biometrics) of the IEEE Biometrics Compendium journal.