Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4411524, 12 pages
https://doi.org/10.1155/2021/4411524

Research Article

A Weakly Supervised Academic Search Model Based on Knowledge-Enhanced Feature Representation

Mingying Xu, Junping Du, Feifei Kou, Meiyu Liang, Xin Xu, and Jiaxin Yang

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Junping Du; junpingdu@126.com

Received 11 May 2021; Revised 30 July 2021; Accepted 11 August 2021; Published 22 September 2021

Academic Editor: Bo Rong

Internet of Things search has great potential applications with the rapid development of Internet of Things technology. Combining Internet of Things technology and academic search to build academic search framework based on Internet of Things is an effective solution to realize massive academic resource search. Recently, the academic big data has been characterized by a large number of types and spanning many fields. The traditional web search technology is no longer suitable for the search environment of academic big data. Thus, this paper designs academic search framework based on Internet of Things Technology. In order to alleviate the pressure of the cloud server processing massive academic big data, the edge server is introduced to clean and remove the redundancy of the data to form a clean data for further analysis and processing by the cloud server. Edge computing network effectively makes up for the deficiency of cloud computing in the conditions of distributed and high concurrent access, reduces long-distance data transmission, and improves the quality of network user experience. For Academic Search, this paper proposes a novel weakly supervised academic search model based on knowledge-enhanced feature representation. The proposed model can relieve high cost of acquisition of manually labeled data by obtaining a lot of pseudolabeled data and consider word-level interactive matching and sentence-level semantic matching for more accurate matching in the process of academic search. The experimental result on academic datasets demonstrate that the performance of the proposed model is much better than that of the existing methods.

1. Introduction

Internet of Things technology has contributed to the realization of many Internet of Things applications that benefit the whole world and constantly changes people's way of life. Internet of Things search service is one of the most important services provided by the Internet of Things. It can efficiently and accurately obtain information to meet the needs of users from massive, heterogeneous, and dynamic Internet of Things data. Currently, academic big data presents the characteristics of huge quantity, various types, and spanning multiple fields. Different from network data in the general sense, academic big data includes paper patents, scientific research equipment, experts and scholars, scientific research teams, national key laboratories, major scientific research infrastructure and large scientific research instruments, academic video scientific research reports, and other scientific and technological entities. Academic big data presents the characteristics of massive, multisource, heterogeneous, and related. Traditional web search technology is no longer suitable for academic big data search environment. Internet of Things search can solve the search problem of academic big data. Internet of Things search is that users send search query requests to the network system, the network system exchanges information with the physical world, and then returns the search object and its location, status, and other information to users. Academic big data search based on Internet of Things technology transfers the massive academic big data to remote cloud server for analysis and calculation, which promotes the interaction between users and intelligent system, improves the search efficiency and realizes automation of academic search. Academic search should not only ensure the search efficiency but also ensure the search accuracy. Therefore, efficient and accurate search is urgently needed.

However, on the one hand, there are several key issues to be considered in the way that massive academic big data is analyzed and calculated by remote cloud server. First of all, multisource, heterogeneous massive big data has data redundancy, noise, data missing, and other problems. If it is transmitted to the remote cloud processing server, it will increase the burden of the cloud server. Secondly, a large number of data directly communicate with the cloud server, which occupies the bandwidth resources and seriously affects the transmission rate. At the same time, accurate academic resource search is another problem to be solved. Many current academic search studies pay little attention on specific search algorithms [1, 2]. There are only few studies on specific academic search task. Explicit Semantic Ranking [3] (ESR) defines academic retrieval as entity set retrieval, which represents the query and each document using knowledge graph embedding, and uses manually labeled training data to train a supervised academic search model. SetRank [4] models the relationships between entities through the type of entities while representing entities and proposes an unsupervised academic search framework. The retrieval performance of ESR and SetRank based on entity set retrieval has been improved to some extent, but there is still much room for improvement. On the one hand, ESR does not consider the relationship between entities. On the other hand, ESR only considers entity level matching, but does not consider deep semantic matching between query and text. What is more, ESR is a supervised academic retrieval model. It can show strong effectiveness [5] when large-scale manual-labeled data of document relevance are available. But manually labeled academic data usually requires domain expert knowledge, which is time-consuming, labor-intensive, and difficult to obtain. It is a serious bottleneck [6] for supervised academic retrieval. Moreover, although SetRank models interentity relationships using entity types, it still cannot model complex relationships between entities. In many cases, a user who submits such a query as “a new ranking method that leverages knowledge graph embedding” will expect to know how “ranking method” associated with “knowledge graph embedding.” The distinguishing characteristic of such queries is that they reflect user’s needs to find documents containing interentity relationships. As the results of the survey in [4], such queries are common in academic search scenarios.

In response to the above-mentioned problems, this paper designs academic search framework in technology Internet of Things. In order to alleviate the pressure of the cloud server processing massive academic big data, the edge server is introduced to clean and remove the redundancy of the data to form a clean data for further analysis and processing by the cloud server. The edge computing network effectively makes up for the deficiency of cloud computing in the conditions of distributed and high concurrent access, reduces the long-distance data transmission process, has faster processing and response speed and lower computing and storage costs, and greatly improves the quality of network user experience. For academic search, this paper proposes a weakly supervised academic search model based on knowledge-enhanced feature representation. In this paper, we refer to both entities and words as features. Specifically, we first employ a scientific information organization tool SCIIE [7] to extract entities and relationships from scientific literatures. In order to obtain the feature representation that can express the entity and the relationship between entities, language model is trained with structured knowledge added as supervision signals. Based on the learned feature vector, we propose a weakly supervised academic retrieval model. In academic search, we should not only pay attention to the semantic matching of words in the query and document but also consider interactive matching between the features in the query and the document. On the other hand, a new weakly supervised method suitable for academic search is employed to obtain a large amount of pseudolabeled training data. Then, the academic search model is trained on labeled training data and pseudolabeled training data based on the knowledge-enhanced feature representation. We conduct extensive experiments on academic data sets. The experimental results demonstrate the effectiveness of our model in improving retrieval performance. We summarize the main novelties and contributions as follows:

(1) An academic search framework based on Internet of Things technology is designed

(2) A novel weakly supervised academic search model based on the knowledge-enhanced feature representation is proposed to improve academic search performance

(3) In the process of academic search, not only the word-level interactive matching but also the sentence-level semantic matching between query and document are considered to realize the accurate matching between query and document

(4) Extensive experiments on the academic datasets prove that our proposed model is significantly better than the state-of-the-art search methods

The rest of this paper is organized as follows: Section 2 discusses related work, and Section 3 describes knowledge-enhanced feature representation for weakly supervised academic search. Section 4 reports experimental results and analysis, and we summarize this article in Section 5.

2. Related Work

2.1. IOT Search. Internet of Things search service is that users send search requests to the network system, the network system exchanges information with the physical world, and then returns information such as the location and status of the search object to users, which is directly driven by users. The existing Internet of Things search technology mainly includes location-based search, content-based search, and heterogeneous search. Location-based search mainly searches the content associated with location. The location information may be expressed as geographic coordinates or may be a logical location, such as a distance from another device [8]. Content-based search is based on the data content collected by a specific target sensor. First, the Internet of Things search engine analyzes the content and maps the corresponding index. Then, when querying, the search engine uses the corresponding index to pair the query with the content and returns the sensor information. Heterogeneous search technology mainly includes semantic or ontology-based search and resource retrieval. In particular, ontology represents concepts, types, and relationships in different fields [9]. The integration of semantic and ontology mechanism can help the system to build domain, task, and method combination search system. On the other hand, all data and IOT devices can abstract resources and provide different services.

2.2. Knowledge-Aware Representation. Representation learning [10] is a core issue in the information retrieval field. Distributed representation methods such as word2vec [11] and glove [12] have been successfully applied in the field of information retrieval [13]. In recent years, the emergence of large-scale knowledge graphs has promoted the development of knowledge-aware representation [14]. The knowledge graph contains rich knowledge [15]. Knowledge-aware representation brings in rich semantic information from the knowledge graph, which significantly improves the effectiveness of the search algorithms [16] and provides new opportunities for better understanding queries and documents [17, 18]. For instance, Xiong et al. introduce a bag-of-entities model, which successfully improves the retrieval accuracy by representing queries and documents using their entity annotations [19]. Hadas et al. design an entity-based language model which uniformly marks the individual features in the text and the term sequence recognized as entities by the entity linking tool as entities and effectively used for document retrieval [20]. A word-entity framework [21] is proposed that combines the bag-of-words and the entities linked to the knowledge graph to optimize information retrieval. Liu et al. employ Entity-Duet Neural Ranking [22], which introduces the knowledge graph into the neural search system. This model proposes a knowledge representation method combining words and entities to represent queries and documents and significantly improves search performance.

2.3. Weakly Supervised Neural Information Retrieval. With the great success of deep neural networks, several deep neural network-based retrieval models have been proposed, such as CDSSM [23], MatchPramid [24], DRMM [25], K-NRM [26], and CONV-KNRM [27]. These methods focus on either matching the whole document or word level interaction. There is no balance between the two aspects. What is more, when large-scale relevance training signals are available, the neural network retrieval models have shown strong effectiveness. However, collecting manually labeled data is time-consuming and labor-intensive [28]. Therefore, insufficient labeled training data has become the main obstacle that affects the performance of neural network-based retrieval models [29]. Weak supervision [30, 31] is an effective way to overcome this limitation. Levy et al. use the generated pseudolabeled data to train the neural network retrieval model, thereby solving the problem of difficulty in obtaining labeled data [32]. One can obtain clicked data of users as weakly supervised training data [33–35]. However, the randomness and subjectivity of user's click behavior caused the inconsistency between user's click and the real relevance label, which made the training set mixed with noise. Dehghani et al. propose a ranking-based method [36] to train a neural retrieval model with the help of an existing retrieval model such as BM25, which assumes that documents with higher scores are more relevant to the query than documents with lower scores. Recently, content-based methods [37, 38] are proposed to generate a set of weak pseudoqueries and related documents. MacAvaney et al. employ filters to eliminate training samples that cannot be converted well into relevance scores [37]. A training sample filtering technique based on heuristics and a new type of supervised filtering [38] is developed to remove those samples which are far away from the domains.

3. Weakly Supervised Academic Search Based on Knowledge-Enhanced Feature Representation in Technology Internet of Things

3.1. System Model. In this part, we introduce the architecture of academic search system in academic Internet of Things. As shown in Figure 1, the system is mainly divided into four layers, data acquisition layer, edge server layer, cloud server layer, and academic search layer.

3.1.1. Data Acquisition. The data acquisition layer mainly obtains academic big data resources, including papers and patents, scientific research teams, scientific and technological talents, key laboratories, instruments, and equipment.

3.1.2. Edge Server. Edge server adopts edge computing model, which is a distributed network model. Because of its high autonomy, each edge server is located in a specific area and connected to a specific device or data source in that area. The edge server is used to clean the academic resources, remove the redundancy, and submit them to the cloud server for further analysis and processing.

3.1.3. Cloud Server. The cloud server layer is composed of a large number of dedicated servers with strong computing power, storage capacity, and stable connection, which can process a large number of data and complex computing in a very short time. Therefore, the most important advantage of the cloud server layer is the aggregation, mining, analysis [39, 40], and storage of massive data, which are beyond the processing capacity of the edge server. The cloud server layer is mainly used to deal with more complex and huge computing intensive data and tasks, such as the training of academic search model based on deep learning.

3.1.4. Academic Search Layer. According to the query submitted by users, it aims to search a group of academic Internet of Things resources, including academic resources and

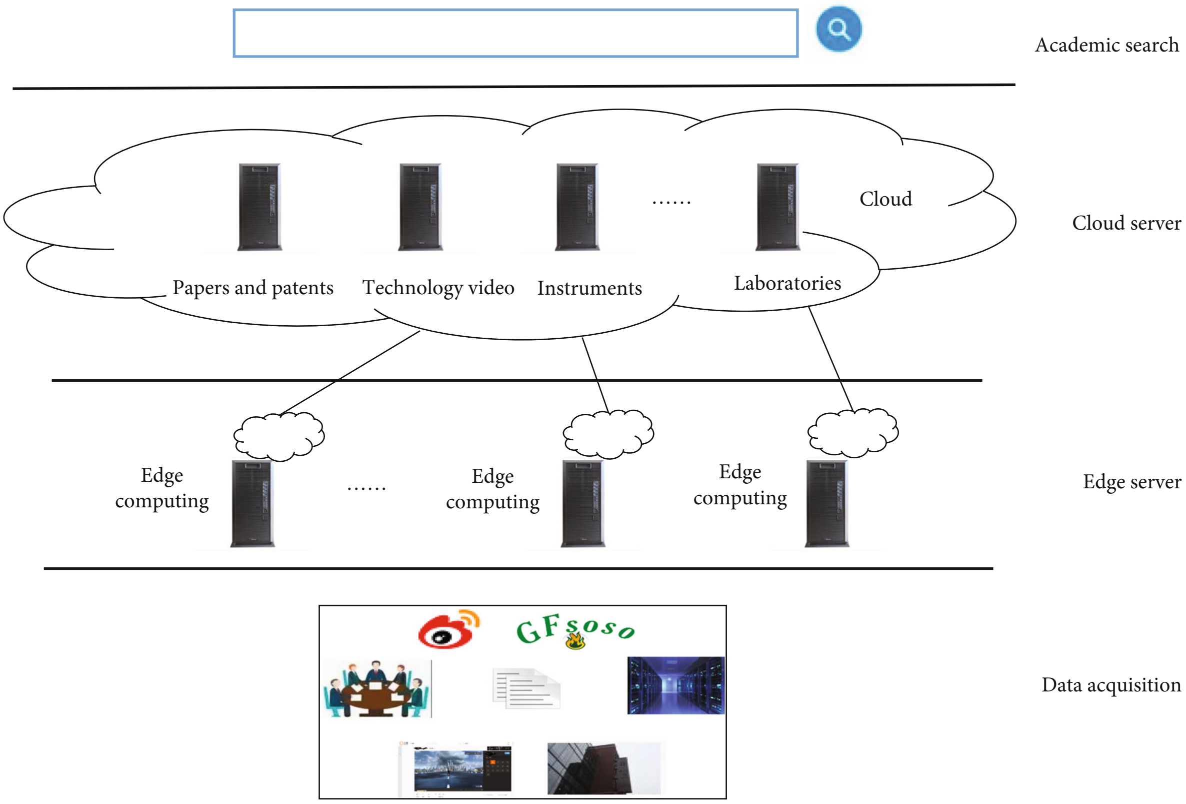


Figure 1: Academic Internet of Things framework.

equipment. The Internet of Things search engine responds to the query and returns academic resources (such as key laboratories and equipment), academic data (such as papers and patents).

The academic search process is shown in Figure 2 as follows.

3.2. Weakly Supervised Academic Search Based on Knowledge-Enhanced Representation. Weakly supervised academic search model is shown in Figure 3, including two main processes: knowledge-enhanced feature learning and weakly supervised academic search.

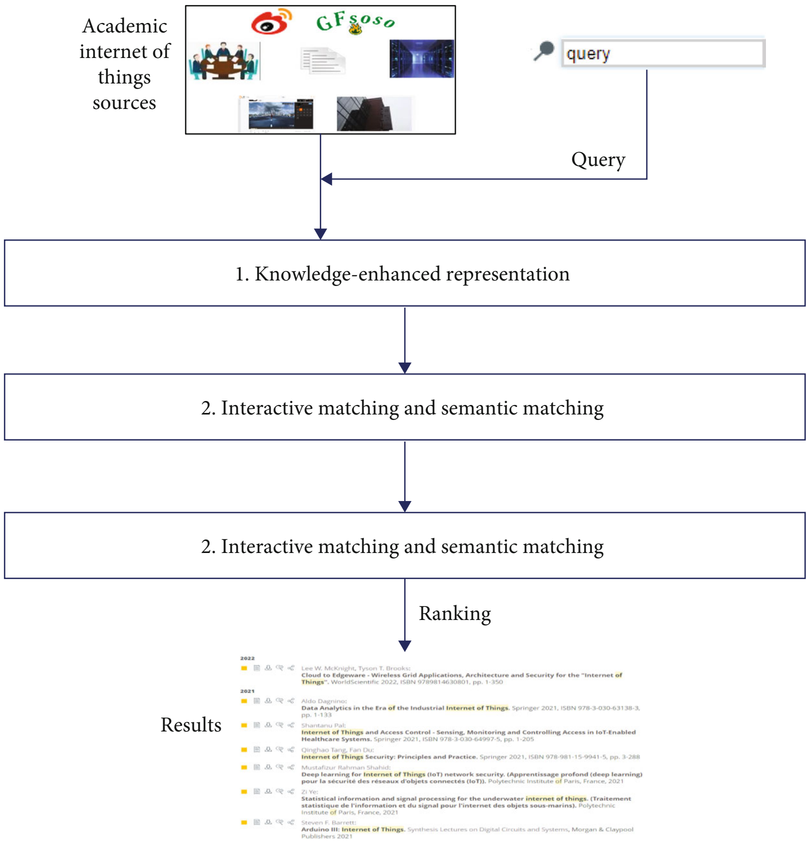
For the knowledge-enhanced feature learning, it models entities and their relationships by adding the knowledge extracted from scientific literatures to train word2vec to express rich semantics. The training process makes the learned features not only depend on the cooccurrence information of features in context but also learn the complex relationship between features.

Weakly supervised academic search model can be trained on large-scale data with the help of weakly supervision strategy to improve the search performance. In the process of academic search, the representation-based matching model and the interaction-based matching model are deployed to the matching module. The model focuses on not only the accurate matching between words but also the semantic matching between query and document to improve the matching accuracy.

3.2.1. Knowledge-Enhanced Feature Representation Learning. In this paper, we propose a knowledge-enhanced feature representation learning method by using word2vec based on skip-gram. The proposed representation learning method is able to express richer semantic information with the knowledge graph embedding model TransE [41]. This paper uses the scientific information extraction tool SCIIE to extract the entities and the relationships in the academic literature. SCIIE defines 6 relationships including "Used-for," "Feature-of," "Hyponym-of," "Part-of," "Compare," and "Conjunction." It can extract entity and interentity relationships in scientific papers and organize them into structured knowledge $T(t_i, r, t_j)$. Here, t_i and t_j are features extracted from the scientific literature, and r is relationship between t_i and t_j. This knowledge is added when training word2vec, so that the learning process of feature vectors is based on not only the cooccurrence information of the context but also the relationship between features, thereby improving the quality of semantic expression. Here, we refer to word and entity extracted from the training corpus as features. The objective function of the knowledge-enhanced feature representation learning model is as follows:

$$L = \sum_i \log p(t_c \mid t_i) + \sum_r \log p\left(t_j \mid t_i + r\right), \tag{1}$$

where t_i is a feature of the entire corpus and term$_c$ is the


Academic internet of things sources
GF soso
query
Query
1. Knowledge-enhanced representation
2. Interactive matching and semantic matching
2. Interactive matching and semantic matching
Ranking
Results

2022

Lee W. McKnight, Tyson T. Brooks:
Cloud to Edgeware - Wireless Grid Applications, Architecture and Security for the "Internet of Things". WorldScientific 2022, ISBN 9789814630801, pp. 1-350

2021

Aldo Dagnino:
Data Analytics in the Era of the Industrial Internet of Things. Springer 2021, ISBN 978-3-030-63138-3, pp. 1-133

Shantanu Pal:
Internet of Things and Access Control - Sensing, Monitoring and Controlling Access in IoT-Enabled Healthcare Systems. Springer 2021, ISBN 978-3-030-64997-5, pp. 1-205

Qinghao Tang, Fan Du:
Internet of Things Security: Principles and Practice. Springer 2021, ISBN 978-981-15-9941-5, pp. 3-288

Mustafizur Rahman Shahid:
Deep learning for Internet of Things (IoT) network security. (Apprentissage profond (deep learning) pour la sécurité des réseaux d'objets connectés (IoT)). Polytechnic Institute of Paris, France, 2021

Zi Ye:
Statistical information and signal processing for the underwater internet of things. (Traitement statistique de l'information et du signal pour l'internet des objets sous-marins). Polytechnic Institute of Paris, France, 2021

Steven F. Barrett:
Arduino III: Internet of Things. Synthesis Lectures on Digital Circuits and Systems, Morgan & Claypool Publishers 2021

Figure 2: Academic search process.

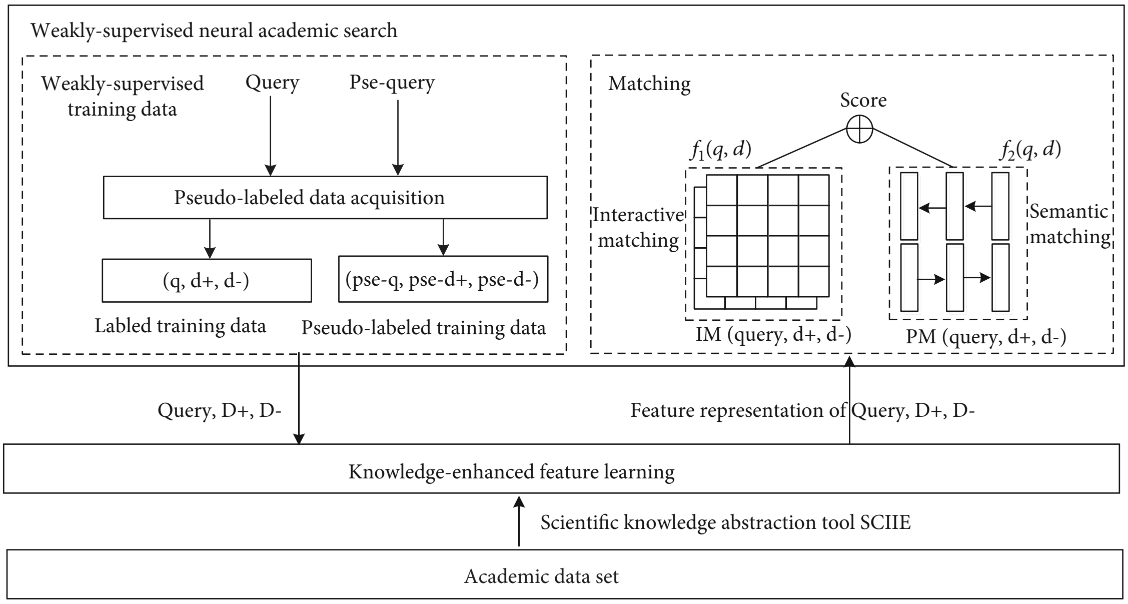

Weakly-supervised neural academic search
Weakly-supervised training data
Query
Pse-query
Pseudo-labeled data acquisition
(q, d+, d-)
(pse-q, pse-d+, pse-d-)
Labled training data
Pseudo-labeled training data
Matching
Score
$f_1(q, d)$
$f_2(q, d)$
Interactive matching
Semantic matching
IM (query, d+, d-)
PM (query, d+, d-)
Query, D+, D-
Feature representation of Query, D+, D-
Knowledge-enhanced feature learning
Scientific knowledge abstraction tool SCIIE
Academic data set

Figure 3: Weakly supervised academic search model.

context of t_i. r is the relationship between t_i and t_j. The left side of Equation (1) is the optimization loss of word2vec based on skip-gram, and the right side is the optimization loss of TransE. $p(t_c \mid t_i)$ is the probability that t_c be predicted as the context of term$_i$. $p(t_j \mid t_i + r)$ is the probability that $t_i + r$ equals to t_j. They can be approximately calculated by negative sampling. Given t_i and r, t_c and t_j are specifically positive samples of word2vec model and TransE model. We use negative sampling to extract negative samples of t_i.

The optimization loss of word2vec based on skip-gram can be calculated as follows:

$$\sum_i \log p(t_c \mid t_i) = \sum_{(t_i,t_c)\in S^+} \log\sigma\left(e_{t_c} \bullet e_{t_i}\right) + \sum_{t_{c'} \epsilon N(t_i)} \log\sigma\left(e_{t_{c'}} \bullet e_{t_i}\right). \tag{2}$$

Here, $(t_i, t_c) \in S^+$ implies that t_i is the context of t_c, S^+ is positive sample sets of t_i. $t_{c'} \epsilon N(t_i)$ implies that $t_{c'}$ is not the context of t_i. $N(t_i)$ is the negative sample set of t_i. e_{t_i}, e_{t_c}, $e_{t_{c'}}$ is the embedding of feature t_i, t_c, $t_{c'}$.

The optimization loss of TransE can be calculated as follows:

$$\begin{aligned} \sum_r \log p\left(t_j \mid t_{\mathrm{i}} + r\right) = & \sum_{\left(t_i,r,t_j\right)\in T^+} \log\sigma\left(e_{t_i+r} \bullet e_{t_j}\right) \\ & + \sum_{\left(t_i,r,t_j\right)\in T^-} \log\left(1 - \sigma\left(e_{t_i+r} \bullet e_{t_j}\right)\right). \end{aligned} \tag{3}$$

$(t_i, r, t_j) \in T^+$implies that $t_i + r$ equals to t_j. T^+ is the positive sample set of (t_i, r). $(t_i, r, t_j) \in T^-$ implies that $t_i + r$ do not equal to t_j. T^- is the negative sample set of (t_i, r).

We use stochastic gradient ascent to update parameters. Finally, feature vectors e_{t_i} can be obtained which express semantic relations among entities and entity relationship.

3.2.2. Weakly Supervised Academic Search. The weakly supervised academic search model is implemented as a pairwise ranking model in Figure 1. Given a query-document pair (q, d_1, d_2), it tries to learn a model that assigns a larger score to document d_1 than document d_2 if d_1 matches to q better. By combining the ranking-based weak supervision method with the content-based weak supervision method, we propose a novel weak supervision method for generating query document pairs.

On the one hand, for a given query and of manually labeled documents, a query-document pair is formed as labeled training data (query, $d+$, $d-$). On the other hand, given a query, top-ranked documents can be recalled from candidate documents using BM25 algorithm. Then, human-labeled documents related to the query are labeled as positive samples, and the documents recalled by BM25 are labeled as negative samples, which forms positive and negative sample pairs. At the same time, considering the title of the paper reflects the most critical content of the paper in the most appropriate and concise terms. And the abstract of the paper contains the most important and necessary information of the paper and is a summary of the main content of the paper. So, we choose the title and corresponding abstract of a paper in the document collection as a positive query-document sample. To sample a negative example of the query, we still use the BM25 algorithm to search the corresponding negative example of this query. Then, the pseudopositive and negative sample pairs are formed, denoted by Pse-training data (Pse-query, Pse-D+, Pse-D-). Finally, weakly supervised training data (Query, $D+$, $D-$) are formed including labeled training data (query, $d+$, $d-$) and pseudolabeled training data (pse-query, pse-d+, pse-d-). The weakly supervised training data are sent to knowledge-enhanced representation layer to map query and document to feature embedding $\{e_t^q\}_{t=1}^m$ and $\{e_t^{id}\}_{t=1}^n$. Then, $\{e_t^q\}_{t=1}^m$ and $\{e_t^{id}\}_{t=1}^n$ are sent to matching layer to calculate matching score1 and score2. Score1 represents the interaction matching score of the word in query and document, and score2 represents the semantic matching score of query and document.

Specifically, KNRM [26] is employed to compute interactive matching score score1. First, the transformation matrix M_{ij} calculates the similarity between the word of query and document.

$$M_{ij} = \cos\left(e_i^q, e_j^d\right). \tag{4}$$

Then, kernels are employed to convert M to query-document matching features $\varnothing(M)$.

$$\begin{aligned} K_K(M_i) &= \sum_j \exp\left(-\frac{\left(M_{ij} - \mu_k\right)^2}{2{\sigma_k}^2}\right), \\ \vec{K}(M_i) &= \{K_1(M_i), \cdots, K_K(M_i)\}, \\ \varnothing(M) &= \sum_{i=1}^n \log\vec{K}(M_i). \end{aligned} \tag{5}$$

$K_K(M_i)$ represents the k^{th} RBF kernel. $\vec{K}(M_i)$ employs K kernels to transform translation matrix into a K-dimensional feature vector. The matching features $\varnothing(M)$ are sent to a fully connected layer to produce the final ranking score $f_1(q, d)$.

$$f_1(q, d) = \tanh\left(W_{s_1}^T \varnothing(\mathrm{M}) + b_{s_1}\right). \tag{6}$$

At the same time, $\{e_t^q\}_{t=1}^m$ and $\{e_t^{id}\}_{t=1}^n$ are sent to a simple Bi-LSTM layer, and we can get the contextual representation. The specific calculation is as follows:

$$\begin{aligned} c_t^q &= \mathrm{BiLSTM}_q\left(c_{t-1}^q, e_i^q\right), \\ c_t^{id} &= \mathrm{BiLSTM}_q\left(c_{t-1}^{id}, e_i^{id}\right). \end{aligned} \tag{7}$$

Then, self-attention is employed to distinguish the importance of different words in the query and sentence of document when calculating the representation of query and document.

$$\begin{aligned} a_t^q &= \frac{\exp\left(W_q^T\left(\tanh\left(W_q \bullet c_t^q\right)\right)\right.}{\sum_{i=1}^{m}\exp\left(W_q^T\left(\tanh\left(W_q \bullet c_t^q\right)\right)\right.}\Bigg), \\ \mathrm{pre}^q &= \sum_{j=1}^{m} a_t^q c_j^q, \\ a_t^{id} &= \frac{\exp\left(W_d^T\left(\tanh\left(W_d \bullet c_t^{id}\right)\right)\right.}{\sum_{j=1}^{n}\exp\left(W_d^T\left(\tanh\left(W_d \bullet c_t^{id}\right)\right)\right.}\Bigg), \\ \mathrm{pre}^{id} &= \sum_{t=1}^{n} a_t^{id} c_t^{id}, \\ \mathrm{pre}^d &= \frac{\sum_{t=1}^{l}\mathrm{pre}^{td}}{l}. \end{aligned} \tag{8}$$

Finally, the similarity score $f_2(q, d)$ of query and document is calculated through a full connection layer.

$$f_2(q, d) = \tanh\left(W_{s_2}^T\left(\mathrm{pre}^q \odot \mathrm{pre}^d\right) + b_{s_2}\right). \tag{9}$$

Given a pair of weakly supervised training samples, we use the hinge loss [42] of information retrieval as the loss function, which is calculated as follows:

$$\begin{aligned} \mathrm{loss} = \sum_{\left(q, d^+, d^-\right) \in \{\mathrm{Query}, D^+, D^-\}} & \max\left(0, 1 - \left(f1\left(q, d^+\right) + f2\left(q, d^+\right)\right)\right. \\ & + \left(f1\left(q, d^-\right) + f2\left(q, d^-\right)\right)), \end{aligned} \tag{10}$$

where $f1(q, d)$ represents the interactive matching score between the query and the document. $f2(q, d)$ represents the semantic matching score between the query and the document to balance human-judged relevance labels and BM25 model scores. The model is fine-tuned using human-labeled relevance judgments after the parameters of the network is pretrained using the weakly supervised data.

4. Experiments Settings

This section describes our experimental dataset, metrics, baselines, and other implementation details.

4.1. Dataset. The experimental datasets come from Semantic Scholar (http://corpus. http://semanticscholar.org/) provided by Xiong et al., including 170,983 candidate document set (https://alleninstitute.org/) with 100 queries. Manual relevance judgement (a 5-level scale, including 4, 3, 2, 1, 0, 4 is the most relevant and 0 is not relevant) are available (http://boston.lti.cs.cmu.edu/appendices/WWW2016/) on the dataset, where both the relevant and irrelevant documents are labeled for each query. In this work, we mainly use the title and abstract of the paper as in [4]. The reasons are as follows: first, the title and abstract of the paper can refine the main points of the paper and summarize the main content of the original text. Second, the original text of the paper is not easy to obtain. Third, the original paper has a large amount of data and slow processing speed. The statistics of experimental datasets is shown in Table 1. Ultimately, we convert this data to a unified format and use this data as an analog for real academic IoT search.

4.2. Baselines and Metrics. We compare information retrieval methods that are popular in the field of information retrieval, including K-NRM [26], Conv-KNRM [27], MatchPyramid [24], DRMM [25], MV-LSTM [43], and ArcII [44].

KNRM: It first computes cosine similarity between query word and document words using a translation layer, and then, it uses a feed-forward network to perform kernel pooling to compute the relevance score between query and document.

Conv-KNRM: Conv-KNRM improves KNRM by using CNN filters to model *n*-gram soft matches of queries and documents to capture *n*-gram such as phrases, concepts, or entities existed in the queries and documents.

MatchPyramid: It computes pair-wise dot product between query and document word vectors to compute an interaction matrix. It then passes this matrix through CNN layers with dynamic pooling to compute the similarity score.

DRMM: It firstly maps the local interaction matrix of query and document into a fixed length matching histogram. Then, it uses forward matching network to learn hierarchical matching features. These features are calculated by term gate network to get the global relevance score of query and document.

MV-LSTM: It uses the word embeddings obtained by passing the sentences through a Bi-LSTM and then computes an interaction vector using cosine similarity or a bilinear operation. It finally passes the interaction vector through a feed-forward network to compute the similarity score.

ArcII: ArcII first computes the interaction feature vector between query and document using CNN layers. It then computes the score for the query-document interaction vector using feed-forward network.

In academic search, the quality of top-ranked results is critical to improve user satisfaction, so the top rankings of Precision and NDCG are particularly important. In order to evaluate the efficiency of academic retrieval model, we report four standard evaluation indicators: MAP, MRR, Precision of top 5, 10, 15, 20 documents retrieved (P@5, P@10, P@15, P@20), NDCG of top 5, 10, 15, 20 documents retrieved (N@5, N@10, N@15, N@20).

4.3. Experiments Setting and Training Details

4.3.1. Knowledge-Enhanced Representation. First, we employ SCIIE to extract the entities and relationships between entities contained in each scientific document. A total of 12,326,751 triples are formed. The detailed knowledge extraction results are shown in Figure 4. From the figure, we can see the proportion of the extracted relationships between entities. We use word2vec based on skip-gram and TransE to train the knowledge-enhanced feature representation, which is optimized by negative sampling. We set the embedding dimension to be 300, and the window size to be 5.

4.3.2. Weakly Supervised Training Data Preparation. First, for each query in the given dataset, the manually labeled documents are formed into positive and negative sample

Table 1: The statistics of academic datasets.

Details	Number
Number of queries	100
Number of papers	170,983
Sum of title length	1,200,641
Sum of abstract length	43,241,159

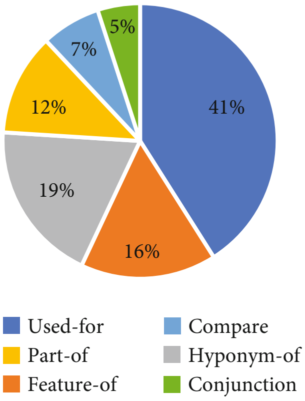


Figure 4: Proportion of relationship between entities in the structured knowledge.

pairs according to the label level. Then, we use the BM25 algorithm to recall 500 documents with higher scores from the candidate documents. In these recalled 500 documents, the documents which have been manually labeled are marked as positive samples, and other documents are marked as negative samples to form positive and negative sample pairs. Then, we use the title of a paper in the document collection as a pseudoquery, and its corresponding document as a positive example. Then, we use BM25 algorithm to search for the corresponding negative examples of this pseudoquery. Finally, a weakly supervised training set is formed, including original labeled data (query, positive samples, negative samples) and prelabeled data (prequery, prepositive samples, and prenegative samples). The model is pretrained on weakly supervised training data and fine-tuned on human-labeled training data.

4.3.3. Effectiveness Verification Settings of the Proposed Model. Our model is abbreviated as KER-IPM-WS. It includes three modules named as knowledge-enhanced representation learning module (KER), matching module (ISM, including interactive matching IM and semantic matching SM), and weakly supervised training (WS). In order to verify the effectiveness of each module, we designed six model variants as follows:

KER-IM: Academic search is implemented based on knowledge-enhanced feature representation, the academic text matching of which is based on interactive matching.

KER-SM: Academic search is implemented based on knowledge-enhanced feature representation; the academic text matching of which is based on interactive matching and semantic matching.

KER-ISM: Academic search is implemented based on knowledge-enhanced feature representation; the academic text matching of which is based on semantic matching.

KER-IM-WS: Weakly supervised academic search is implemented based on knowledge-enhanced feature representation; the academic text matching of which is based on interactive matching.

KER-SM-WS: Weakly supervised academic search is implemented based on knowledge-enhanced feature representation; the academic text matching of which is based on semantic matching.

KER-ISM-WS: Weakly supervised academic search is implemented based on knowledge-enhanced feature representation; the academic text matching of which is based on interactive matching and semantic matching.

We implement our model using TensorFlow. For the neural retrieval models CDSSM, KNRM, and Conv-KNRM, we use an open-source implementation MatchZoo (https://github.com/NTMC-Community/MatchZoo). For KNRM and Conv-KNRM, we set the number of bins to 11. We apply gradient descent algorithm and Adam as our optimizer for training the ranking model. And we use dropout technology to prevent overfitting. We set the batch size to 64 and select the learning rate from $[1e-1, 1e-2, 1e-3, 1e-4]$. The training epoch number is set to 20. We use 5-fold crossvalidation to validate our model. We randomly split all training data into five equal partitions. In each fold, three partitions are used for training, one for validation and one for testing.

5. Result Analysis and Use Case

5.1. Result Analysis. In this section, we first verify the effectiveness of knowledge-enhanced representation by comparing with the retrieval methods in baselines. Then, we further compare the academic retrieval performance of different module of KER-ISM-WS and analyze the impact on retrieval performance of different module of KER-ISM-WS.

5.2. Retrieval Performance of Different Retrieval Model. Tables 2 and 3, respectively, show the retrieval performance of weakly supervised academic search based on knowledge-enhanced feature representation and the baselines on NDCG@{5, 10, 15, 20} and Precision@{5, 10, 15, 20}. Table 4 shows the retrieval performance of the proposed weakly supervised academic search based on knowledge-enhanced feature representation and the baselines on MAP and MRR.

It can be seen from Tables 2, 3, and 5, the retrieval performance of our model exceeds the baselines. Experimental results show that our method is effective for academic search. It is worth noting that ConV-KNRM employs CNN filters to model n-gram of queries and documents to identify phrases, concepts, or entities. In addition to our model, ConV-KNRM achieves the best retrieval performance in the baselines. These findings indicate that mining phrases, concepts, or entities existed in academic texts is helpful for improving academic retrieval performance.

Table 2: Performance comparison on NDCG of different retrieval models.

Model	N@5	N@10	N@15	N@20
KNRM	0.4160	0.4807	0.5406	0.6077
ConvKNRM	0.4437	0.5258	0.5891	0.6535
MatchPyramid	0.3655	0.4475	0.5056	0.5727
DRMM	0.3935	0.4536	0.5411	0.5996
MV-LSTM	0.3244	0.4013	0.4822	0.5525
ArcII	0.4256	0.4768	0.5587	0.6094
KER-IPM-WS	0.5842	0.6009	0.6396	0.6653

Table 3: Performance comparison on precision of different retrieval models.

Model	P@5	P@10	P@15	P@20
KNRM	0.5241	0.5103	0.4987	0.4653
ConvKNRM	0.5452	0.5327	0.5048	0.4852
MatchPyramid	0.4517	0.4310	0.4321	0.3978
DRMM	0.4647	0.4414	0.4187	0.3831
MV-LSTM	0.4207	0.4034	0..3895	0.3527
ArcII	0.4655	0.4264	0.4207	0.3862
KER-IPM-WS	0.6509	0.6232	0.5521	0.5248

Table 4: Performance comparison on NDCG of different module of KER-ISM-WS.

Model	N@5	N@10	N@15	N@20
KER-IM	0.4753	0.5075	0.5672	0.6157
KER-SM	0.4427	0.4869	0.5406	0.5993
KER-ISM	0.5293	0.5693	0.5919	0.6269
KER-IM-WS	0.5228	0.5452	0.5816	0.6436
KER-SM-WS	0.4865	0.5133	0.5791	0.6255
KER-ISM-WS	0.5842	0.6009	0.6396	0.6653

Table 5: Performance comparison on MAP and MRR of different retrieval models.

Model	MAP	MRR
KNRM	0.4173	0.4929
ConvKNRM	0.4521	0.5402
MatchPyramid	0.3849	0.4730
DRMM	0.4158	0.4775
MV-LSTM	0.3437	0.4208
ArcII	0.4485	0.5086
KER-IPM-WS	0.5569	0.6427

5.3. The Impact on Retrieval Performance of Different Module of KER-ISM-WS. This section shows the impact of different module of KER-ISM-WS on retrieval performance Precision, NDCG, MAP, and MRR. The results are shown in Tables 4 and 6.

Table 6: Performance comparison on precision of different module of KER-ISM-WS.

Model	P@5	P@10	P@15	P@20
KER-IM	0.5733	0.5409	0.4961	0.4628
KER-SM	0.5439	0.5028	0.4874	0.4379
KER-ISM	0.5931	0.5217	0.5037	0.4867
KER-IM-WS	0.6031	0.5478	0.4856	0.4701
KER-SM-WS	0.5834	0.5254	0.4713	0.4443
KER-ISM-WS	0.6509	0.6232	0.5521	0.5248

Table 7: Performance comparison on MAP and MRR of different module of KER-ISM-WS.

Model	MAP	MRR
KER-IM	0.4764	0.5921
KER-SM	0.4491	0.5317
KER-ISM	0.5252	0.6437
KER-IM-WS	0.5121	0.5958
KER-SM-WS	0.4964	0.5838
KER-ISM-WS	0.5569	0.6427

5.3.1. The Impact on Retrieval Performance of Knowledge-Enhanced Feature Representation. KER_IM is academic search based on knowledge-enhanced feature representation; the academic text matching of which is based on KNRM. Therefore, we can compare it with KNRM to verify the effectiveness of the knowledge-enhanced representation. By comparing the retrieval performance of KNRM in Tables 2, 3, and 5 and KER_IM in Tables 4, 6, and 7, we can conclude that the proposed knowledge-enhanced feature representation method can improve retrieval performance.

5.3.2. The Impact on Retrieval Performance of Matching Based on Interactive Matching and Semantic Matching. We can compare KER_IM, KER_SM, and KER_ISM to verify the effectiveness of the matching based on interactive matching and semantic matching. From Tables 4, 6, and 7, the results indicate that the retrieval performance of the model KER_ISM is better than KER_IM and KER_SM. The reason is that KER_ISM considers both word interaction matching and sentence semantic matching to guide more accurate search.

5.3.3. The Impact on Retrieval Performance of Weakly Supervised Training. We can compare KER_IM and KER_IM_WS, KER_SM and KER_SM_WS, and KER_ISM and KER-ISM-WS to verify the effectiveness of the weakly supervised training. From Tables 5–7, the results indicate that the retrieval performance of the models KER-KER_IM-WS, KER_SM_WS, and KER_ISM_WS is better than that of KER-IM, KER-SM, and KER-ISM. The possible reason is that a large amount of weakly supervised training data can guide the model to learn better parameters.

Table 8: Search results of query “object detection”.

KER-PM-WS	Paper title	Human label	Entity relation
1	Rapid **object detection** using a boosted cascade of simple features	4	Cascade of simple features **USED-FOR object detection**
2	Class-specific Hough forests for **object detection**	5	Hough forests **USED-FORobject detection**
3	Faster R-CNN: towards real-time **object detection** with region proposal networks	3	R-CNN **USED-FOR object detection**
4	Learning rich features from RGB-D images for **object detection** and segmentation	2	Learning rich features **USED-FORobject detection** RGB-D images **USED-FOR** learning rich features RGB-D images **USED-FORobject detection**
5	Rich feature hierarchies for accurate **object detection** and semantic segmentation	3	Rich feature hierarchies **USED-FOR object detection**

Query: object detection.

5.4. Use Case—Academic Search. In this section, we use the proposed KER_ISM_WS model to report the academic search results. A real use case for literature retrieval is as follows in Table 8. The input query is object detection, which reflects the information needs of users.

Table 8 shows the retrieval results of our model and the manually labeled level for these retrieval results. Entities in the titles of the retrieved papers are highlighted in green. Our method also gives the entities that exist in the title of the paper and the relationships between them. Entities are marked as green; relationships are marked as red.

From the search results and manually labeled tags, we can see that our model has achieved accurate matching. At the same time, we infer from the manually annotated data that the query is trying to find documents that have a certain relationship with the “object detection.” The search results show that our model can retrieve articles which meet the needs. Moreover, all the items in the search results can express the entities related to the query and the relationship information between the entities. This further illustrates the effectiveness of the proposed representation learning model. It not only models the entities but also models the semantic relationships that exist between entities, which helps to improve the performance of the retrieval model.

6. Conclusion

We design academic Internet of Things search framework based on Internet of Things technology. In order to alleviate the pressure of the cloud server processing massive academic big data, the edge server is introduced to clean and remove the redundancy of the data to form a clean data for further analysis and processing by the cloud server. For academic search, we propose a novel knowledge-enhanced feature representation learning method which can express rich semantics in texts of academic search field. Aiming at the “data hungry” property of deep neural academic retrieval methods, we propose a weakly supervised academic search model based on the knowledge-enhanced representation, which relieves high cost of acquisition of manually labeled data by obtaining a lot of pseudolabeled data in the process of academic search. In the process of text matching for academic retrieval, the proposed model considers both the word-level interactive matching and the sentence-level semantic matching to improve matching accuracy of relevance. Experiment results on real academic search datasets show that the proposed model is effective and greatly improves the academic search performance.

Data Availability

The datasets we use in this paper can be downloaded from http://boston.lti.cs.cmu.edu/appendices/WWW2016/.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFB1402600), National Natural Science Foundation of China (NSFC) under grant (No. 61772083, No. 61802028, and No. 61877006).

References

[1] J. Zhang and J. Tang, “Name disambiguation in AMiner,” *Science China-information sciences*, vol. 64, no. 4, 2021.

[2] G. Abramo, C. A. D’Angelo, and F. di Costa, “A gender analysis of top scientists' collaboration behavior: evidence from Italy,” *Scientometrics*, vol. 120, no. 2, pp. 405–418, 2019.

[3] X. Chenyan, P. Russell, and C. Jamie, “Explicit semantic ranking for academic search via knowledge graph embedding,” in *Proceedings of the 26th International Conference on World Wide Web*, Perth, Australia, 2017.

[4] J. Shen, X. Jinfeng, X. He, S. Jingbo, S. Saurabh, and H. Jiawei, "Entity set search of scientific literature: an unsupervised ranking approach," in *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ann Arbor, MI, USA, 2018.

[5] S. Marchesin, A. Purpura, and G. Silvello, "Focal elements of neural information retrieval models. An outlook through a reproducibility study," *Information Processing and Management*, vol. 57, no. 6, article 102109, 2020.

[6] K. Zhang, C. Xiong, Z. Liu, and Z. Liu, "Selective weak supervision for neural information retrieval," in *Proceedings of The Web Conference*, Taipei, Taiwan, 2020.

[7] L. Luan, L. He, O. Mari, and H. Hannaneh, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[8] S. Liang and C. Y. Huang, "GeoCENS: a geospatial cyberinfrastructure for the world-wide sensor web," *Sensors*, vol. 13, no. 10, pp. 13402–13424, 2013.

[9] S. Pattar, R. Buyya, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "Searching for the IoT resources: fundamentals, requirements, comprehensive review, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2101–2132, 2018.

[10] L. Chen, D. Chen, F. Yang, and J. Sun, "A deep multi-task representation learning method for time series classification and retrieval," *Information Sciences*, vol. 555, pp. 17–32, 2021.

[11] M. Tomas, S. Ilya, C. Kai, C. Greg, and D. Jeffrey, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, 2013.

[12] P. Jeffrey, S. Richard, and D. Manning Christopher, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[13] M. Bhaskar and C. Nick, "An introduction to neural information retrieval," *Foundations and Trends in Information Retrieval*, vol. 13, no. 1, pp. 1–126, 2019.

[14] D. Hongliang, T. Siliang, and F. Wu, "Entity mention aware document representation," *Information Science*, vol. 430, pp. 216–227, 2018.

[15] O. P. Ghiasnezhad, K. Wang, and Z. Wang, "An embedding-based approach to rule learning in knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1348–1359, 2021.

[16] J. Huang, H. Wang, W. Zhang, and T. Liu, "Multi-task learning for entity recommendation and document ranking in web search," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 5, pp. 1–24, 2020.

[17] T. Komamizu, "Random walk-based entity representation learning and re-ranking for entity search," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2989–3013, 2020.

[18] C. Xiong, Z. Liu, J. Callan, and T. Liu, "Towards better text understanding and retrieval through kernel entity salience modeling," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, MI, USA, 2018.

[19] C. Xiong, J. Callan, and T. Liu, "Bag-of-entities representation for ranking," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, Newark, DE, USA, 2016.

[20] H. Raviv, O. Kurland, and D. Carmel, "Document retrieval using entity-based language models," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, 2016.

[21] C. Xiong, J. Callan, and T. Liu, "Word-entity duet representations for document ranking," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, 2017.

[22] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Entity-duet neural ranking: understanding the role of knowledge graph semantics in neural information retrieval," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[23] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the 23th International Conference on World Wide Web*, Seoul, Korea, 2014.

[24] P. Liang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, 2016.

[25] J. Guo, Y. Fan, Q. Ai, and W. Croft, "A deep relevance matching model for ad-hoc retrieval," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, Indianapolis, IN, USA, 2016.

[26] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, 2017.

[27] Z. Dai, C. Xiong, J. Callan, and Z. Liu, "Convolutional neural networks for soft-matching n-grams in ad-hoc search," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Marina Del Rey, CA, USA, 2018.

[28] Y. Zheng, Y. Liu, Z. Fan et al., "Investigating weak supervision in deep ranking," *Data and Information Management*, vol. 3, no. 3, pp. 155–164, 2019.

[29] H. Zamani and W. Croft, "On the theory of weak supervision for information retrieval," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, Ann Arbor, MI, USA, 2018.

[30] R. Levy, B. Bogin, S. Gretz, R. Aharonov, and N. Slonim, "Towards an argumentative content search engine using weak supervision," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 2018.

[31] J. Han, Y. Yang, D. Zhang, D. Huang, D. Xu, and F. de la Torre, "Weakly-supervised learning of category-specific 3D object shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1423–1437, 2021.

[32] H. Zamani and W. Croft, "Towards theoretical understanding of weak supervision for information retrieval," in *Proceedings of the ACM SIGIR Workshop on Learning from Limited or Noisy Data for Information Retrieval*, Ann Arbor, MI, USA, 2018.

[33] B. Li, P. Cheng, and L. Jia, "Joint learning from labeled and unlabeled data for information retrieval," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 2018.

[34] X. Ling, W. Deng, C. Gu, H. Zhou, C. Li, and F. Sun, “Model ensemble for click prediction in bing search ads,” in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, Perth, Australia, 2017.

[35] L. Xue, Z. Jianjin, L. Zhipeng et al., “Learning fast matching models from weak annotations,” in *Proceedings of the World Wide Web Conference*, San Francisco, CA, USA, 2019.

[36] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, and W. Croft, “Neural ranking models with weak supervision,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, 2017.

[37] S. Mac Avaney, K. Hui, and A. Yates, “An approach for weakly-supervised deep information retrieval,” 2017, http://arxiv.org/abs/1707.00189.

[38] S. Mac Avaney, A. Yates, K. Hui, and O. Frieder, “Content-based weak supervision for ad-hoc re-ranking,” in *Proceedings of the 42th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France, 2019.

[39] X. Sui, M. Li, Y. Ying et al., “Aerolysin nanopore identification of single nucleotides using the AdaBoost model,” *Journal of Analysis and Testing*, vol. 3, no. 2, pp. 134–139, 2019.

[40] T. H. Fereja, F. Du, C. Wang, D. Snizhko, Y. Guan, and G. Xu, “Electrochemiluminescence imaging techniques for analysis and visualizing,” *Journal of Analysis and Testing*, vol. 4, no. 2, pp. 76–91, 2020.

[41] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, 2013.

[42] J. Guo, Y. Fan, L. Pang et al., “A deep look into neural ranking models for information retrieval,” *Information Processing and Management*, vol. 57, no. 6, p. 102067, 2020.

[43] S. Wan, Y. Lan, J. Guo, J. Xu, P. Liang, and X. Cheng, “A deep architecture for semantic matching with multiple positional sentence representations,” in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, 2016.

[44] B. Hu, Z. Lu, H. Li, and Q. Chen, “Convolutional neural network architectures for matching natural language sentences,” in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014.