JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 18, No. 1, March 2024, pp. 36-46

TRIO: An Entity Retrieval Method Using Entity Embedding and Topic Modeling

Hyejin Park
Korea Electronics Technology Institute, Seoul, Korea
hyejinpark@keti.re.kr

Daecheol Woo
Department of Computer Science, Yonsei University, Seoul, Korea
dwoo@yonsei.ac.kr

Shinwoo Park
Department of Artificial Intelligence, Yonsei University, Seoul, Korea
pshkhh@yonsei.ac.kr

Kyungwon Kim*
Korea Electronics Technology Institute, Seoul, Korea
kwkim@keti.re.kr

Abstract
Current entity search models predominantly rely on term frequency and semantic similarity, often failing to fully exploit the information in the knowledge graphs. This limitation leads to the neglect of entities that could be highly relevant to the user's query topics. To overcome these challenges and enhance entity retrieval, we introduce TRIO (Term, topic, and neural-based entity Retrieval Interpolate methOd), an entity retrieval method that employs multiple search perspectives for more relevant outcomes. TRIO stands out by seamlessly integrating three distinct search perspectives: term frequency, semantic similarity, and topic similarity. This integration is executed in a simple and effective manner, allowing TRIO to capture entities across multiple dimensions, resulting in comprehensive and accurate search results. Our experiments on the standard DBpedia-Entity V2 test collection demonstrate a substantial enhancement in the search performance of the baseline model. On average, TRIO improves NDCG and MAP performance by 12.401%, and 27.342%, respectively, compared to the best-performing baseline model.

Category: Information Retrieval / Web

Keywords: Information retrieval; Knowledge graph; Entity embedding, Knowledge service; Deep learning

Open Access **http://dx.doi.org/10.5626/JCSE.2024.18.1.36** **http://jcse.kiise.org**

Received 10 October 2023; **Accepted** 14 March 2024
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

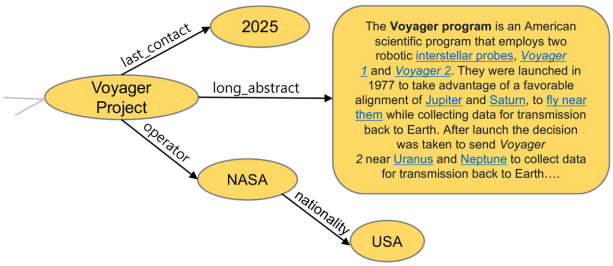
I. INTRODUCTION

The increasing use of knowledge graphs (KGs) in various real-world applications has highlighted the significance of entity search [1–3]. Entity search assesses the relevance of entities based on information needs expressed in natural language or keyword queries. The existing approach falls short in its searching capacity. First, the term-based approach [4–7] measures the relative significance of each entity by calculating the term-matching score between entity and queries. This method does not use semantic word meanings, relying solely on alphabetical matching, thereby resulting in a vocabulary gap. Second, neural-based approaches [8–11] include encoding KG triples or local paths of KG into a low-dimensional embedding space. Although these neural methods alleviate the vocabulary gap by generating embedding values that align with semantic similarities between words, they are not devoid of shortcomings. Neural-based models consider the local context of the words, facilitating searches limited to semantically relevant entities. Semantic word similarity can inadvertently omit numerous entities that necessitate a broader scope for relevant search results

KGs are extracted from diverse sources, leading to mixed-labeled data representations within KGs. For example, the facts about "Voyager Project" in Fig. 1(a) are represented in either structured label or unstructured free text label formats. Both types of data contain valuable information within KGs, warranting the consideration of both data formats within the retrieval model. Nonetheless, many neural-based approaches dismiss or intentionally exclude this facet given that unstructured text labels are often excessively lengthy, posing challenges for models to handle, or due to the perception that the information within these labels holds minimal significance [12, 13].

On the other hand, entity retrieval within a single perspective (term-based or neural-based) of KGs is limited to searching entities observable only within their own dimensions, and this limitation is compounded by the constraints of mixed-labeled data representations, which may hinder full utilization of KGs' potential. To overcome this limitation research efforts have been made to enhance the entity retrieval performance by combining neural-based and term-based approaches [8], or by incorporating topic modeling into the description of unstructured node entities within KGs, thereby redefining entity descriptions [13]. While research amalgamates search perspectives to enhance retrieval performance, it still grapples with the task of capturing query intent entirely.

In light of these insights, we introduce TRIO (Term, topic, and neural-based entity Retrieval Interpolate methOd), a simple yet powerful entity retrieval method. TRIO capitalizes on the strengths of three distinct approaches: term-based method, neural-based method, and topic modeling-based method. Utilizing topic modeling enhances



(a)

Q: Wonders of the ancient world

: Highly relevant : Weakly relevant : Irrelevant

	<Term-based Model>	<Term+Neural based Model>	<TRIO>
1	Seven Wonders of the Ancient World	Seven Wonders of the Ancient World	Pyramid of Giza
2	Seven Ancient Wonders	Wonders of the World	Colossus of Rhodes
3	Wonders of the World	Seven Ancient Wonders	Temple of Artemis
4	Falconcity of Wonders	*Primeval civilization*	Seven Wonders of the Ancient World
5	The Seven Fabulous Wonders	*Giza Necropolis*	*Giza Necropolis*

(b)

Fig. 1. (a) Facts can exist in structured and unstructured formats. For instance, the fact "The operator of Voyager Project is NASA" can be represented as a structured triple (Voyager Project, operator, NASA). Additionally, facts related to the long abstract for the Voyager Project are depicted as unstructured text. (b) Retrieved entities are ranked by relevance score. Term-based model results are underlined, entities retrievable by neural-based searches are in italics, and highly relevant TRIO-retrieved entities are highlighted in yellow.

the understanding of search intent by extracting meaningful topics from the unstructured text content [12, 14–18]. Term-based search strengthens relevance through precise term matches, while neural-based search considers semantic similarity, uncovering broader related entities and increasing search flexibility. By overlapping the search results of each dimension, the comprehensive search approach improves results and better aligns with user intent. By integrating the three perspectives, TRIO not only expands the range of retrievable entities beyond what previous models could achieve but also harnesses the heterogeneous data present in KGs. We conducted an experiment that explored entity search results for the query "Wonders of the ancient world," depicted in Fig. 1(b). The term-based method retrieves exact term matches like "Seven Wonders of the Ancient World," but is limited to literal matches. A blend of term-based and neural-based models identifies primeval civilization by recognizing the semantic tie between ancient and primeval, though still falling short of complete query intent comprehension. Introducing topic similarity via TRIO yields triumphant results, capturing precise user intent with the *Pyramid of Giza*, *the Colossus of Rhodes*,

and the *Temple of Artemis*. Thus, TRIO's multi-perspective enhances entity retrieval, unveiling a deeper understanding of user queries. Our main contributions are summarized as:

- A pragmatic method that combines term frequency, semantic, and topic similarity in entity search models.
- An overview of previously hidden entities, utilizing KGs' heterogeneity to enhance query intent understanding.
- An adaptable mechanism that enhances neural-based search models' performance by incorporating our approach as an additional layer.

II. RELATED WORK

A. Term-based Retrieval

Conventional entity search methods primarily involve retrieving entities through term matching between queries and KG nodes or by treating KGs as structured documents for direct term matching between queries and KG content. These techniques rely on term alignment to facilitate entity retrieval based on correspondences between queries and knowledge graph entities. Treating KGs as structured documents also offers an alternative means for conducting direct term-based entity searches between queries and KG materials. Several knowledge-base (graph) based retrieval methods are proposed. User query retrieval systems leverage knowledge bases with distinct attributes [4]. The studies similar to the method of retrieving knowledge by transforming queries are [5–7, 19]. Studies are using the KGs as an instance of structured documents and using it for generating entity representations. Fielded sequential dependence model (FSDM) [20] improves retrieval performance by defining an entity representation schema consisting of five layers to capture meaning and relations between entities. Studies similar to using KGs to generate entity representations are [21, 22]. This approach lacks consideration for the semantic meanings of words, relying solely on alphabetical matching, resulting in a limitation in capturing the full extent of vocabulary comprehension.

B. Neural-based Retrieval

In the embedding-based retrieval research, a neural-based approach emerges as a powerful solution to overcome the limitations of term-based retrieval. By vectorizing KGs into low-dimensional spaces, this approach addresses the constraints posed by diverse frameworks commonly utilized within KGs. EMB [23] introduces a neural network that embeds KG entities and relations into continuous vectors, overcoming KGs' disparate frameworks. TransE [24] simplifies EMB complexity by embedding triplets ($\mathbf{e}_h$, $\mathbf{r}$, $\mathbf{e}_t$) in the same vector space, with $\mathbf{e}_h + \mathbf{r}$ close to $\mathbf{e}_t$, enabling simpler model training and embedding of large KGs. Moreover, recent studies have explored efficient neural network models, leveraging document and query encoding through dual encoders [25], and investigating neural network architectures for efficient retrieval by combining dense and sparse retrieval models into hybrid frameworks [26]. Additionally, a novel method has been proposed to enhance the generalization ability of dense retrieval models [27]. In contrast, our method enhances entity retrieval by integrating various search perspectives effectively, primarily focusing on an approach that integrates diverse search perspectives: term frequency, semantic similarity, and topic similarity.

Diverse pre-trained language models are explored for entity retrieval research. Elas4RDF [9] improves entity retrieval in open-domain question answering (QA) systems via keyword-based methods, including answer type prediction, SPARQL-based entity enrichment, and pre-trained neural models. Similarly, an efficient entity search model [28] employs a pre-trained bidirectional encoder representations from transformers (BERT) model [29]. Query by webpage [11] incorporates new feature extraction modules, including a text feature extraction module (TEM), enhancing think tank quality using term and topic-based features. There are studies combining other entity retrieval methods. Utilizing DeepWalk for graph embedding and Word2Vec for word embedding, knowledge graph entity and word embedding for retrieval (KEWER) [8] embeds KG relations and entities and enhances search with BM25F. EntityLDA [13] proposes an entity retrieval methodology to generate an entity representation with a richer representation by utilizing subject modeling for a non-unified node representation of the knowledge graph and to better interpret the intent of a user query. However, Fig. 1(b), a single-perspective entity retrieval, still presents room for enhancing the understanding of user query intent, which can be addressed through the integration of multi-dimensional search approaches.

III. METHODOLOGY

Term-based methods rely on term-matching scores but have vocabulary gaps and semantic limitations. In contrast, neural-based methods address vocabulary gaps using low-dimensional embeddings, but ignore heterogeneous data and focuses only on local context, potentially missing query intent in search results. This study aims to enhance query intent comprehension and create a streamlined entity search approach. We integrate topic similarity with term-based and semantic methods, harmonizing retrieval outcomes to improve precision.

Our TRIO comprises term-based, neural-based, and topic-modeling-based search components. These are collectively detailed in Sections III-A, III-B, and III-C.

A. Entity Search with Term-based Approach

BM25F [30, 31], an extension of the Best Match 25 (BM25) algorithm, measures document relevance by evaluating the frequency of terms within documents using the bag-of-words concept. Incorporating BM25F as a term-based search mechanism presents the advantage of precisely identifying crucial fields within KG documents that have been restructured into multi-field formats. Notably, BM25F has demonstrated exceptional performance in comparison to other term-based retrieval methods, as evidenced in its superior results within the DBpedia-Entity V2 test collection [32]. This high-performing technique has also been integrated into the prior state-of-the-art entity search approach, KEWER [8]. By employing the multi-fielded KG node entity representation schema proposed in FSDM [20], BM25F retrieves relevant documents (entities) for each query and returns the results.

B. Entity Search with Neural-based Approach

In recent years, the intersection of deep learning and information retrieval (IR) through artificial neural networks has garnered considerable attention. This symbiotic relationship has led to the emergence of various forms of neural-based IR research [33]. To comprehensively measure TRIO's effectiveness across a spectrum of neural models, we systematically integrate distinct pre-existing neural-based retrieval models into the TRIO framework. This strategic approach not only shows how TRIO synergizes with diverse neural-based retrieval strategies but also substantiates its potential to enhance retrieval performance across different methodologies. Furthermore, we leverage the neural component of TRIO as a benchmark, establishing it as the baseline model to substantiate and validate the amplified retrieval capabilities that it brings forth (for more details, refer to Section IV-B).

C. Entity Search with Topic Modeling-based Approach

Despite being a traditional topic model, latent Dirichlet allocation (LDA) [34, 35] offers a straightforward and intuitive approach that can be easily applied, justifying its usage for various purposes. Rather than training models on the entire KGs, LDA focuses on topic extraction, enhancing tasks like document retrieval. This approach is effective for managing large KG collections. LDA's attributes can be harnessed to effectively leverage unstructured information within KGs. By utilizing this capability, it becomes feasible to extend the perspective of entity retrieval into a topic-based search. In our approach, we employ LDA to discern latent topics from FSDM entity documents containing unstructured information about KG node entities. This endeavor enables the extraction of topics associated with node entities within the KG, which can subsequently be utilized as rich representations for these entities. This strategic application of LDA plays a pivotal role in enhancing the performance of entity retrieval tasks, particularly in the context of heterogeneous data. We retrieve the related entity to the query according to the relevance scoring formula in [36].

The basic idea is exploiting the probability that the word w that makes up the user query q appears in each topic t and the probability of what topics appear in the entity document d corresponding to entity e to calculate the topic relevance score between the query and the entity as:

$$LDA(q,e) = \sum_{w}^{|q|} \sum P(w \mid t) P(t \mid d) \qquad (1)$$

For each query, the LDA-based search component generates a list of the retrieved entities in the order of relevance score.

D. General Framework

As shown in Fig. 2, the general framework of TRIO has three major components: 1) BM25F for term-based search; 2) arbitrary neural-based search model for semantic-based search; and 3) LDA for topic-based search. Each component independently performs entity searches based on its unique perspective and inherent characteristics, aiming to identify pertinent entities relevant to the query. Subsequently, the results obtained from these three components are carefully integrated. This fusion of diverse search dimensions enhances the approach, providing a comprehensive and enhanced entity retrieval capability.

We seamlessly merge the search outcomes through the following three steps:

1. Each entity search component measures the relevance

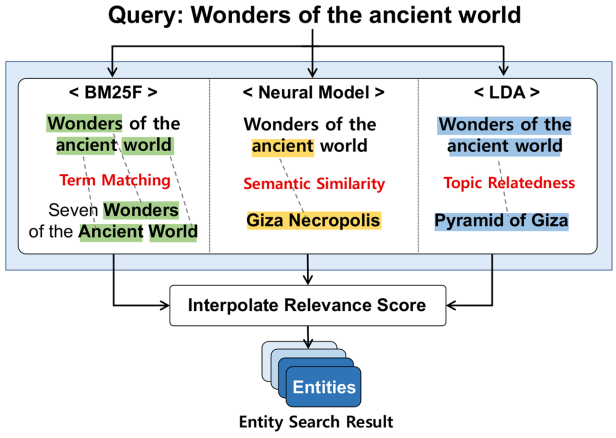


Fig. 2. Structure of TRIO. For the given query, TRIO retrieves the entities combining three search perspectives. After retrieving, it interpolates and returns a rank of entities.

of entity e with respect to the user query q.
2. The search models yield a list of entities ranked by their relevance scores.
3. The resulting entity lists are blended using linear interpolation.

The influence of each search component on the search results is quantified by the weights α, β, and γ (refer to Section IV-C for specifics), denoted as follows:

$$\begin{aligned}\text{TRIO}(q, e) = \alpha \cdot \text{BM25F}(q, e) + \\ \beta \cdot \text{Neural Model}(q, e) + \\ \gamma \cdot \text{LDA}(q, e),\end{aligned} \quad (2)$$

where $\alpha + \beta + \gamma = 1$.

IV. EXPERIMENTAL SETUP

A. Dataset

Knowledge graph: We employ the DBpedia 2015-10 [37] as the KG for our entity retrieval task. The DBpedia community's meticulous extraction techniques have rendered this KG highly reliable and well-defined.

Entity documents: In our entity retrieval process, we utilize the entity representation schema introduced by FSDM [20], corresponding to the entities present in the DBpedia KG nodes. These entity documents contain comprehensive descriptions, including unstructured information about the entity. Drawing inspiration from the KEWER study, we employ Galago [38] to index the entity representation across five distinct fields (names, categories, similar entity names, attributes, and related entity names), thereby enhancing the capture of inter-entity meanings and relationships.

Test dataset: To assess the performance of TRIO, we employ the DBpedia-Entity V2 [32] as our evaluation dataset. DBpedia-Entity V2 serves as a standardized benchmark for evaluating entity retrieval models based on the DBpedia KG. This dataset encompasses 496 queries across four distinct categories (named entity queries, keyword queries, natural language queries, and search particular lists) as detailed in Table 1. For determining query relevance, the DBpedia-Entity V2 collection provides an entity ranking list comprising DBpedia URIs along with their corresponding relevance scores.

Table 1. DBpedia-Entity V2 dataset overview

Category	Description	Count
QALD2	Natural language questions	140
INEX-LD	IR-style keyword queries	154
SemSearch ES	Named entity queries	156
ListSearch	Queries for specific entity lists	46

B. Baselines

JOINTLY(J) [39] serves as a fundamental baseline for the previously established state-of-the-art entity retrieval approach, KEWER [8]. JOINTLY employs an embedding technique that unifies entities, relations, and words within a shared embedding space by leveraging entity description-based alignment. Our approach adopts the KEWER implementation of JOINTLY. KEWER offers multiple versions of JOINTLY, depending on whether they exploit entity linking or surface form. We select the version that demonstrates the best performance among the available alternatives as our baseline model.

Elas4RDF$_{Ro}$ (E). Among the components of the Elas4RDF keyword search pipeline, Elas4RDF$_{Ro}$ (E) stands out as one of its integral tasks [9]. Elas4RDF encompasses diverse tasks including answer type prediction, entity enrichment through SPARQL, and the extraction of answers via pre-trained neural model RoBERTa [40]. In our exploration, we focus on Elas4RDF$_{Ro}$, a specific facet of Elas4RDF that capitalizes on answer entities derived from RoBERTa. This model merges the outcomes of the Elas4RDF search engine with answer entities, ranging from 1 to 10, obtained from RoBERTa. In our assessment, we pinpoint the optimal case of search performance, marked by the amalgamation of 10 answer entities, to serve as our baseline model.

KEWER(K) [8] employs a low-dimensional embedding approach to unify entities and words, capturing both local structure and structural components within the KG. Utilizing graph random walks over the KG, KEWER(K) considers its local structure and components, then employs the negative skip-gram-based Word2Vec model to embed these random walks. In practice, KEWER retrieves and assesses entities for query relevance by computing cosine similarity based on entity embedding values. To address the modest performance of KEWER's embedding module in search, it is complemented with BM25F. The integration of BM25F and KEWER's embedding module significantly amplifies search effectiveness, yielding an ad-hoc entity search approach.

C. Hyperparameter Settings

The coefficients of Eq. (2) are fine-tuned through five-fold cross-validation. We determine optimal hyper-parameters α, β, and γ by iteratively adjusting each parameter from zero to one in increments of 0.002, aiming to maximize the average NDCG@10 value within each fold of the training set from DBpedia-Entity V2. When we utilize JOINTLY as the neural component within TRIO, the optimal model weight parameters average to $\alpha = 0.135$, $\beta = 0.244$, and $\gamma = 0.613$. In the case of Elas4RDF$_{Ro}$, the optimal parameters are $\alpha = 0.054$, $\beta = 0.611$, and $\gamma = 0.335$. For the KEWER scenario, the optimal weight parameters are $\alpha = 0.113$, $\beta = 0.509$, and $\gamma = 0.379$.

D. Implementation Details

While generating an entity document following the FSDM approach [20], we utilize the latest version of Galago to index the entity representation into four fields: names, attributes, categories, and related entity names. Although the FSDM suggests five fields, including similarity names in addition to the four aforementioned fields, our experimentation revealed slightly better performance when employing only these four fields.

V. RESULTS AND ANALYSIS

A. Evaluation Metrics

To assess the effectiveness of the retrieval method, we employ two standard evaluation metrics: normalized discovered cumulative gain (NDCG) [41] and mean average precision (MAP). For NDCG, we evaluate the search performance at different cutoffs: top 10 (NDCG@10), top 30 (NDCG@30), top 50 (NDCG@50), and top 100 (NDCG@100).

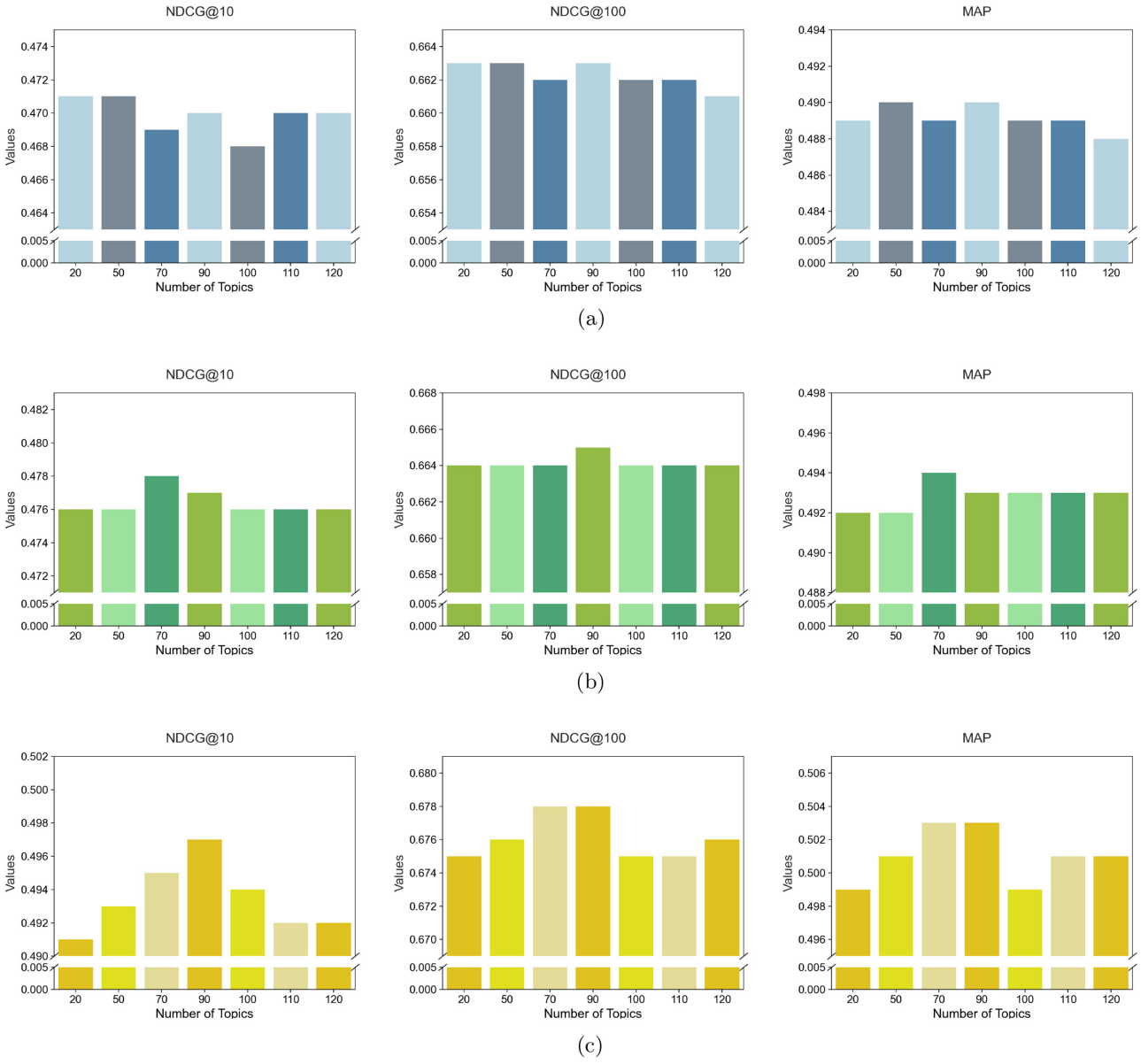


Fig. 3. Search performance by the number of topics. $TRIO_{NM}$ refers to an entity search method using the neural model NM: (a) $TRIO_J$, (b) $TRIO_E$, and (c) $TRIO_K$.

B. Effect of Number of Topics on Entity Search Performance

To determine the optimal number of topics that maximizes retrieval performance, we analyze how TRIO's performance varies with different numbers of topics. Fig. 3 demonstrates that TRIO consistently exhibits favorable search performance when the number of topics is below 100, irrespective of the neural components used. Notably, 90 topics yields the highest average performance. In the case of TRIO*J*, it excels with 20 and 50 topics for NDCG@10. However, it also displays strong retrieval outcomes with 90 topics for NDCG@100 and MAP. For $TRIO_E$ and $TRIO_K$, we observe that overall search performance diminishes, except at 70 and 90 topics.

Choosing very few topics within TRIO leads to the omission of vital information within entity documents, essential for effective retrieval, consequently impacting search quality. Conversely, an excess of topics introduces superfluous noise, hampering accurate retrieval. Thus, the selection of an optimal number of topics significantly influences the search performance. We designate $t = 90$ as the number of topics for the LDA model in subsequent experiments in Section V-C and V-D, given its superior entity search performance, as depicted in Fig. 3.

C. Overall Performance

Table 2 presents the entity search performance achieved by integrating our proposed approach with existing neural-based methods.

In the JOINTLY experiment, JOINTLY embeds only words and geometric features of KG components for search purposes. However, the outcomes of searches relying solely on a single neural perspective are inferior compared to those employing other singular perspective-based approaches such as BM25F and LDA, as depicted in Table 2. This performance gap suggests that JOINTLY struggles to effectively capture the intent behind user queries. Hence, to gain a deeper comprehension of user intent, it becomes imperative to enhance entity retrieval by incorporating term match-based and topic similarity-based viewpoints. Additionally, model weight parameters outlined in Section IV-C reveal that the retrieval accuracy of $TRIO_J$ substantially benefits from the introduction of the LDA-based search. Furthermore, Table 2 illustrates that $TRIO_J$ significantly enhances search performance across various top numbers of retrieved entities, as measured by the NDCG metric, in comparison to utilizing JOINTLY alone. Regarding search accuracy assessed through the MAP metric, $TRIO_J$ improves search precision, surpassing the retrieval accuracy achieved by JOINTLY in isolation. In the case of $Elas4RDF_{Ro}$, it elevates the performance of the Elas4RDF search engine by augmenting the list of entities retrieved from the search engine with answer-entities generated by RoBERTa. Despite attempts to enhance search performance through the incorporation of neural models, $Elas4RDF_{Ro}$ encounters limitations in its entity search capabilities. Nevertheless, it is feasible to further enhance search effectiveness by integrating perspectives related to topics and term matching. As shown in Table 2, $TRIO_E$ surpasses the average retrieval accuracy of $Elas4RDF_{Ro}$ by 39%, demonstrating a notable improvement. Moreover, the retrieval accuracy doubles, underscoring the significant enhancement achieved.

For the KEWER scenario, the model incorporates both the semantic viewpoint and the structural characteristics of KGs, enhancing search performance through the addition of term-based search BM25F. However, the results shown in Fig. 1(b) reveal that KEWER has limitations in comprehending the user's query intent, leaving room for potential search performance improvement. TRIOK addresses KEWER's search vulnerabilities by merging three distinct search perspectives. As indicated in Table 2, $TRIO_K$ showcases a notable average improvement of 12.4% in overall retrieval performance, with retrieval accuracy rising by 27.3%. In summary, our TRIO model

Table 2. Entity search performance according to different neural-based models

	NDCG@10	NDCG@30	NDCG@50	NDCG@100	MAP
BM25F	0.463	0.486	0.509	0.542	0.379
LDA	0.229	0.319	0.405	0.518	0.330
JOINTLY	0.151	0.141	0.143	0.152	0.080
$TRIO_J$	**0.470**	**0.520**	**0.577**	**0.663**	**0.490**
$Elas4RDF_{Ro}$	0.298	0.326	0.354	0.383	0.237
$TRIO_E$	**0.477**	**0.524**	**0.578**	**0.665**	**0.493**
KEWER	0.480	0.497	0.520	0.554	0.395
$TRIO_K$	**0.497**	**0.543**	**0.595**	**0.678**	**0.503**

BM25F and LDA are the single-perspective search models, each of which is used as a term-based and topic-modeling part of TRIO. TRIO-NM refers to an entity search method using the neural model NM as a component. The bold font indicates the best performance in each test.

surpasses other baseline approaches, underscoring the effectiveness of TRIO's three distinct search perspectives for efficient entity retrieval.

D. Ablation Study

To establish the robustness of our approach, we conducted an ablation study to evaluate retrieval performance by removing individual components from TRIO. The results of this study are presented in Table 3. According to our findings, all three components of TRIO—BM25F, a neural-based model, and an LDA-based model—contribute significantly to the overall effectiveness of the TRIO approach.

The least impact on the effectiveness of TRIO is observed when the neural component (TRIO-NM) is removed. Although it might initially appear that the neural-based retrieval method does not play a crucial role in TRIO, it's evident that this component serves a unique purpose in performing semantic-based searches, distinct from term match-based and topic similarity-based searches. The optimal weight parameters for TRIO components assign substantial weights to the three neural-based search models (refer to Section IV-C). When utilizing $\text{Elas4RDF}_{\text{Ro}}$ and KEWER as the neural- based component in TRIO, the corresponding weight values are 0.611 and 0.509, respectively. The weights above 0.5 emphasizes the significant contribution of the neural-based component to the search process.

The removal of the LDA-based component (TRIO-LDA) significantly affects the effectiveness of TRIO. Although eliminating the LDA-based component has minimal impact on NDCG@10, the disparity with the complete TRIO search performance widens as the number of evaluated entities increases, such as in the case of NDCG@100. LDA-based entity search involves extracting topics from entity documents and using these topics to calculate similarity with the query, thereby identifying entities related to the query. Given that LDA-based search operates in a distinct dimension compared to term-matching-based and semantic similarity-based search, it's evident that the LDA-based component greatly influences TRIO's retrieval performance. Moreover, the LDA-based component holds significant weights of 0.613 for TRIO_{J}, and weight values of 0.335 and 0.379 for TRIO_{E} and TRIO_{K} respectively, according to the optimal weight parameter settings for the TRIO components.

Removing the BM25F component (TRIO-BM25F) significantly affects the effectiveness of TRIO. This is due to the fact that certain entities, with high relevance scores, contain words that match those in the query, and BM25F excels at identifying entities that share the same query terms. Moreover, irrespective of the neural components in TRIO, the MAP results show lower search accuracy when any single component is removed, compared to the complete TRIO. This reaffirms that the strength of TRIO lies in the meticulous integration of its three components.

VI. DISCUSSION & FUTURE WORK

Our approach offers a simple yet effective method to boost entity retrieval models. By integrating three distinct search perspectives, we tap into different dimensions of entity retrieval. Term-based search hinges on query term frequency, neural-based search delves into word meanings and relationships, and topic modeling-based search evaluates topic relevance. This fusion creates a comprehensive search model that broadens the scope of entity retrieval possibilities, suggesting exciting avenues for future research.

On the practical side, our work aids researchers in fine-tuning topic modeling-based entity retrieval models. Our experiments reveal the optimal number of topics for effective retrieval using LDA-based topic modeling. We find that the number of topics significantly influences

Table 3. Impact of removing entity search components from TRIO (NM denotes neural model)

	NDCG@10	NDCG@30	NDCG@50	NDCG@100	MAP
TRIO-NM	0.464	0.499	0.527	0.562	0.389
TRIO_{J}	**0.470**	**0.520**	**0.577**	**0.663**	**0.490**
-LDA	0.467	0.489	0.513	0.540	0.384
-BM25F	0.337	0.367	0.389	0.416	0.279
TRIO_{E}	**0.477**	**0.524**	**0.578**	**0.665**	**0.493**
-LDA	0.470	0.494	0.517	0.549	0.385
-BM25F	0.309	0.378	0.423	0.444	0.266
TRIO_{K}	**0.497**	**0.543**	**0.595**	**0.678**	**0.503**
-LDA	0.480	0.497	0.520	0.554	0.395
-BM25F	0.349	0.431	0.508	0.600	0.406

TRIONM - (·) denotes the entity search method without a component (·). The bold font indicates the best performance in each test.

search performance. Striking the right balance is crucial, as too few or too many topics affect the performance. Our findings, specifically the peak performance at 90 topics, streamline the decision-making process for practitioners utilizing topic modeling for entity searches.

While TRIO brings advancements, it has certain limitations. Exploring neural-based topic modeling and refining component combination are two promising directions for future research. By considering neural-based topic modeling and exploring alternative ways to merge components, we aim to push the boundaries of TRIO's search performance even further in our future research

VII. CONCLUSION

While neural-based search models alleviate the vocabulary gap seen in term-based searches, they often struggle with entity retrieval involving topic relevant to the query due to their focus on local contexts. In contrast, KGs capture information from diverse sources, leading to non-uniform data representation in both structured and unstructured forms.

This paper introduced TRIO, an entity retrieval method that blends three search perspectives: term-based, neural-based, and topic modeling-based. TRIO offers a simple yet impactful approach to enhancing search performance, adaptable to various neural-based models. Through a comprehensive set of experiments across neural-based search methods, we demonstrated TRIO's versatility and improved entity search performance compared to baseline methods.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01352, Development of technology for validating the autonomous driving services in perspective of laws and regulations).

REFERENCES

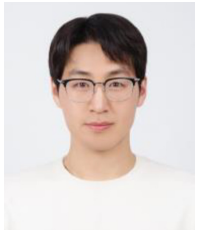
1. X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications*, vol. 141, article no. 112948, 2020. https://doi.org/10.1016/j.eswa.2019.112948
2. J. Dalton and L. Dietz, "Constructing query-specific knowledge bases," in *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC)*, San Francisco, CA, USA, 2013, pp. 55-0. https://doi.org/10.1145/2509558.2509568
3. R. Reinanda, E. Meij, and M. de Rijke, "Knowledge graphs: an information retrieval perspective," *Foundations and Trends in Information Retrieval*, vol. 14, no. 4, pp. 289-444, 2020. https://doi.org/10.1561/1500000063
4. S. Shekarpour, A.C. Ngonga Ngomo, and S. Auer, "Question answering on interlinked data," in *Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 1145-1156. https://doi.org/10.1145/2488388.2488488
5. J. Pound, A. K. Hudek, I. F. Ilyas, and G. Weddell, "Interpreting keyword queries over web knowledge bases," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, USA, 2012, pp. 305-314. https://doi.org/10.1145/2396761.2396803
6. T. Tran, D. M. Herzig, and G. Ladwig, "SemSearchPro: using semantics throughout the search process," *Journal of Web Semantics*, vol. 9, no. 4, pp. 349-364, 2011. https://doi.org/10.1016/j.websem.2011.08.004
7. A. Tonon, G. Demartini, and P. Cudre-Mauroux, "Combining inverted indices and structured search for ad-hoc object retrieval," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, OR, USA, 2012, pp. 125-134. https://doi.org/10.1145/2348283.2348304
8. F. Nikolaev and A. Kotov, "Joint word and entity embeddings for entity retrieval from a knowledge graph," in *Advances in Information Retrieval*. Cham, Switzerland: Springer, 2020, pp. 141-155. https://doi.org/10.1007/978-3-030-45439-5_10
9. C. Nikas, P. Fafalios, and Y. Tzitzikas, "Open domain question answering over knowledge graphs using keyword search, answer type prediction, SPARQL and pre-trained neural models," in *The Semantic Web–ISWC 2021*. Cham, Switzerland: Springer, 2021, pp. 235-251. https://doi.org/10.1007/978-3-030-88361-4_14
10. P. Jafarzadeh and F. Ensan, "A semantic approach to post-retrieval query performance prediction," *Information Processing & Management*, vol. 59, no. 1, article no. 102746, 2022. https://doi.org/10.1016/j.ipm.2021.102746
11. Q. Geng, Z. Chuai, and J. Jin, "Webpage retrieval based on query by example for think tank construction," *Information Processing & Management*, vol. 59, no. 1, article no. 102767, 2022. https://doi.org/10.1016/j.ipm.2021.102767
12. X. Li, J. Mao, W. Ma, Y. Liu, M. Zhang, S. Ma, Z. Wang, and X. He, "Topic-enhanced knowledge-aware retrieval model for diverse relevance estimation," in *Proceedings of the Web Conference 2021*, Ljubljana, Slovenia, 2021, pp. 756-767. https://doi.org/10.1145/3442381.3449943
13. Y. Hong, S. Feng, and Y. Xiao, "EntityLDA: a topic model for entity retrieval on knowledge graph," in *Proceedings of 2020 IEEE International Conference on Knowledge Graph (ICKG)*, Nanjing, China, 2020, pp. 388-395. https://doi.org/10.1109/ICBK50248.2020.00062
14. M. N. Awan and M. O. Beg, "Top-rank: a TopicalPostionRank for extraction and classification of keyphrases in text," *Computer Speech & Language*, vol. 65, article no. 101116, 2021. https://doi.org/10.1016/j.csl.2020.101116

15. A. Curiel, C. Gutierrez-Soto, and J. R. Rojano-Caceres, "An online multi-source summarization algorithm for text readability in topic-based search," *Computer Speech & Language*, vol. 66, article no. 101143, 2021. https://doi.org/10.1016/j.csl.2020.101143
16. S. Miles, L. Yao, W. Meng, C. M. Black, and Z. B. Miled, "Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling," *Information Processing & Management*, vol. 59, no. 3, article no. 102921, 2022. https://doi.org/10.1016/j.ipm.2022.102921
17. F. Jian, J. X. Huang, J. Zhao, T. He, and P. Hu, "A simple enhancement for ad-hoc information retrieval via topic modelling," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy, 2016, pp. 733-736. https://doi.org/10.1145/2911451.2914748
18. M. Mendoza, P. Ormeno, and C. Valle, "Ad-hoc information retrieval based on boosted latent Dirichlet allocated topics," in Proceed*ings of 2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, Santiago, Chile, 2018, pp. 1-7. https://doi.org/10.1109/SCCC.2018.8705252
19. T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-based interpretation of keywords for semantic search," in *The Semantic Web*. Heidelberg, Germany: Springer, 2007, pp. 523-536. https://doi.org/10.1007/978-3-540-76298-0_38
20. N. Zhiltsov, A. Kotov, and F. Nikolaev, "Fielded sequential dependence model for ad-hoc entity retrieval in the web of data," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, 2015, pp. 253-262. https://doi.org/10.1145/2766462.2767756
21. R. Neumayer, K. Balog, and K. Norvag, "On the modeling of entities for ad-hoc entity search in the web of data," in *Advances in Information Retrieval*. Heidelberg, Germany: Springer, 2012, pp. 133-145. https://doi.org/10.1007/978-3-642-28997-2_12
22. F. Nikolaev, A. Kotov, and N. Zhiltsov, "Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy, 2016, pp. 435-444. https://doi.org/10.1145/2911451.2911545
23. A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 301-306, 2011. https://doi.org/10.1609/aaai.v25i1.7917
24. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2787-2795, 2013.
25. Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329-345, 2021. https://doi.org/10.1162/tacl_a_00369
26. D. Lee, S. W. Hwang, K. Lee, S. Choi, and S. Park, "On complementarity objectives for hybrid retrieval," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canadam 2023, pp. 13357-13368. https://doi.org/10.18653/v1/2023.acl-long.746
27. S. Xu, L. Pang, H. Shen, and X. Cheng, "BERM: training the balanced and extractable representation for matching to improve generalization ability of dense retrieval," 2023 [Online]. Available: https://arxiv.org/abs/2305.11052.
28. E. J. Gerritse, F. Hasibi, and A. P. de Vries, "Entity-aware transformers for entity search," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, 2022, pp. 1455-1465. https://doi.org/10.1145/3477495.3531971
29. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2009, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423
30. H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson, "Microsoft Cambridge at TREC 13: Web and Hard Tracks," in *Proceedings of the 13th Text Retrieval Conference (TREC)*, Gaithersburg, Maryland, USA, 2004, pp. 1-7.
31. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333-389, 2009. https://doi.org/10.1561/1500000019
32. 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 *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tokyo. Japan, 2017, pp. 1265-1268. https://doi.org/10.1145/3077136.3080751
33. K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, et al., "Neural information retrieval: at the end of the early years," *Information Retrieval Journal*, vol. 21, pp. 111-182, 2018. https://doi.org/10.1007/s10791-017-9321-y
34. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
35. G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, USA, 2009, pp. 758-759. https://doi.org/10.1145/1571941.1572114
36. X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, USA, 2006, pp. 178-185. https://doi.org/10.1145/1148170.1148204
37. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: a nucleus for a web of open data," in *The Semantic Web*. Heidelberg, Germany: Springer, 2007, pp. 722-735. https://doi.org/10.1007/978-3-540-76298-0_52
38. M. A. Cartright, S. J. Huston, and H. Feild, "Galago: a modular distributed processing and retrieval system," in *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval (OSIR@SIGIR)*, Portland, OR, USA, 2012, pp. 25-31.
39. H. Zhong, J. Zhang, Z. Wang, H. Wan, and Z. Chen, "Aligning knowledge and text embeddings by entity descriptions," in *Proceedings of the 2015 Conference on Empirical Methods*

in Natural Language Processing, Lisbon, Portugal, 2015, pp. 267-272. https://doi.org/10.18653/v1/d15-1031
40. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "Roberta: a robustly optimized BERT pretraining approach," 2019 [Online]. Available: https://arxiv.org/abs/1907.11692.
41. K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422-446, 2002. https://doi.org/10.1145/582415.582418

Hyejin Park https://orcid.org/0009-0005-4847-2116

Hyejin Park received her master's degree in Computer Science from Yonsei University, Seoul, Korea, in 2022. Since then, she has been working as a Researcher at the Korea Electronics Technology Institute in Seoul, Korea. Her research interests focus on (Hyper)Graph Learning and utilizing graphs for information retrieval and recommendation systems.



Daecheol Woo https://orcid.org/0000-0001-6887-5282

Daecheol Woo received master's degree in Computer Science from Yonsei University, Seoul, Korea in 2023. His research interests lie in natural language processing and information retrieval based on knowledge graph.

Shinwoo Park https://orcid.org/0000-0001-8704-5751

Shinwoo Park is currently enrolled in the integrated master's and doctoral program at Yonsei University, Seoul, starting from 2020. His research interests include code AI, with a focus on natural language code search.

Kyungwon Kim https://orcid.org/0000-0001-9537-9225

Kyungwon Kim is received Ph.D. degree in Computer, Information and Communications Engineering from Konkuk University, Seoul, Korea, in March 2018. He has been a Principal Researcher at Korea Electronics Technology Institute in Seoul, Korea, since 2004. His current research interests include the unstructured data analysis and data inference modeling.