Scientometrics (2024) 129:2181–2203
https://doi.org/10.1007/s11192-024-04966-9

Knowledge graph enhanced citation recommendation model for patent examiners

Yonghe Lu[1,2] · Xinyu Tong[1] · Xin Xiong[1] · Hou Zhu[1]

Received: 23 April 2023 / Accepted: 8 February 2024 / Published online: 11 March 2024

Abstract
In the face of a growing volume of patents, patent examiners grapple with prolonged examination cycles, prompting the need for more effective citation recommendations. To address this, we introduce the patent knowledge graph embedded in Bert (PK-Bert) model. This innovation combines a patent knowledge graph with semantic information in an advanced Transformer framework, outperforming conventional common-sense knowledge graph embedding. PK-Bert exhibits substantial improvements, boosting the recall of accurate citation recommendations by 2.15% over the benchmark model Bert and 1.25% over K-Bert with CnDBpedia. Ablation experiments highlight the significance of knowledge graph elements, with the inventor proving most influential, followed by the IPC number and assignee. At the same time, publication time and title information have a minor impact. Moreover, PK-Bert excels when trained with earlier data and evaluated for patents issued post-November 2023. Our study not only advances patent examiner recommendations but also presents an efficient integration method for knowledge graph-enhanced semantic patent characterization.

Keywords Knowledge graph · Patent citation recommendation · Patent examiner citation · Deep learning

Introduction

As the volume of patents continues to escalate, patent examiners confront extensive patent databases, necessitating the meticulous identification of relevant citations to support their examination endeavors (Alcácer & Gittelman, 2006; Kuhn, 2011). A proficient citation recommendation system not only economizes examiners' time but also enhances their comprehension of pertinent prior implementations, thereby elevating the scientific robustness and dependability of patent examinations (Choi et al., 2022a, 2022b). Moreover, precision in citation recommendation plays a pivotal role in fostering knowledge dissemination, thereby stimulating innovation and technological advancement (Färber & Jatowt, 2020).

✉ Hou Zhu
zhuhou3@mail.sysu.edu.cn

[1] School of Information Management, Sun Yat-sen University, Guangzhou, China

[2] School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

Unlike scientific and technical papers that often cite many references to support research backgrounds or theories, patent citations have unique citation motives and purposes (Brooks, 1985; Meyer, 2000). In addition to the patent applicant's self-review of the patent application, patent citations primarily come from the patent examiner's review of technically relevant "comparison documents" of the patented innovation (Lin et al., 2016). In order to enhance the standardization and rigor of patent grants and prevent applicants from intentionally omitting citations of technically similar literature to bolster their chances of patent approval, patent examiners must thoroughly examine all technically analogous literature associated with the proposed patent application (Zhao & Wen, 2017). Therefore, examiner citations reflect knowledge linkage better than applicant citations and are a good indicator of knowledge linkage, and patent citations that are technologically relevant to the text of the pending patent can be obtained through the examiner's evaluation (Chen, 2017; Wada, 2018).

Although models applying various methods for patent citation recommendation (Chen et al., 2020; Yin et al., 2024) are now available, these models still ignore the importance of examiner citations. As an important means of protecting intellectual property, the patent examination process is strict and careful (China National Intellectual Property Office (CNIPO), 2010). In this regard, existing recommendation systems must fully utilize the potential of more reliable patent citation information labeled by professional patent examiners. Patent examiners are responsible for determining the patentability of pending patents, and thus, examiner citation datasets may be more credible and applicable than applicant citations used in previous studies (Choi et al., 2022a, 2022b; Fu et al., 2015; Lu et al., 2020).

Patent citations are established upon technical citation contexts, encompassing factors such as semantic similarity, technical relevance, or legal protection existing between the pending patent and the cited patent (Meyer, 2000). Therefore, to gain a comprehensive understanding of the context of patent citations, it is essential to consider not only the textual data of patents but also to integrate metadata that represents the technical domain. However, prior studies have primarily focused on identifying semantically similar patents (An et al., 2021; Arts et al., 2018; Zhang et al., 2016) or relied solely on textual similarity for citation recommendation (Chen et al., 2020; Yin et al., 2024). These approaches might not have adequately adapted to the diverse citation recommendation needs of patent examiners.

Hence, the primary focus of this paper is to leverage knowledge graphs to enhance the accuracy and efficiency of citation recommendations in the patent examination process. Through the integration of structured information, entities, relationships, and patent-specific semantic details, the objective is to offer patent examiners more contextually relevant and precise recommendations. Our code and data are available at https://github.com/xinyu0610/Citation-Recommendation-Model-for-Patent-Examiners.

The main contributions of this study can be described as follows:

(i) We propose PK-Bert, utilizing a knowledge graph with semantic information and an improved Transformer framework, to identify patent examiner citation relationships.
(ii) We collect crucial patent attributes to construct the Patent-info knowledge graph, comprising a total of 215,089 patent triples.
(iii) We introduce a novel approach to embed patent information and investigate the impact of various attributes on examiner citation recommendations through ablation experiments.

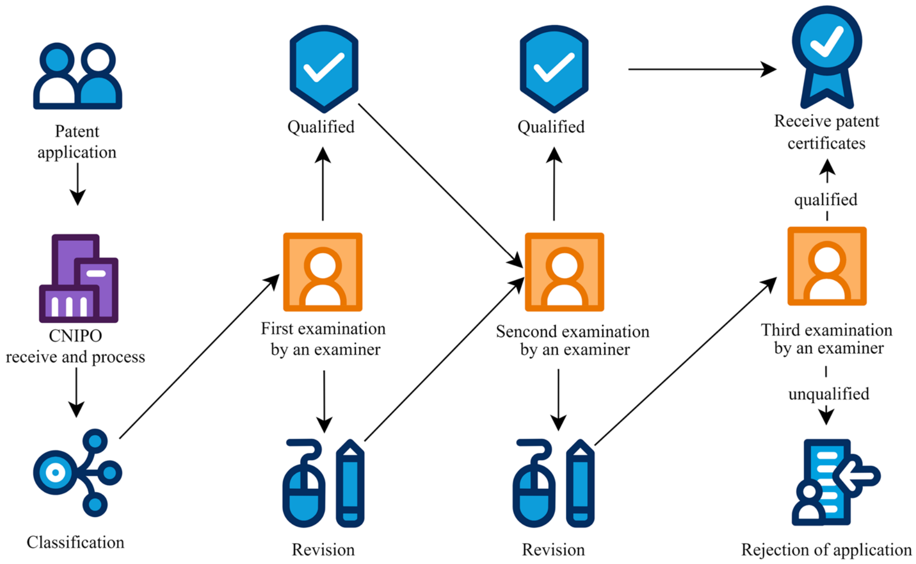

Patent application
CNIPO receive and process
Classification
Qualified
First examination by an examiner
Revision
Qualified
Sencond examination by an examiner
Revision
Receive patent certificates
qualified
Third examination by an examiner
unqualified
Rejection of application

Fig. 1 Simplified patent examination process for Chinese invention patents

The remainder of this manuscript is organized as follows. We review previous literature in "Related works" section and "Methodology" section describes the methods and models used in this study. "Experiment" section shows the data collection, knowledge graph construction, patent examiner citation recommendation experiments, and the ablation study. Finally, the discussion and conclusion are presented in "Conclusion and future works" section.

Related works

In this section, we review and summarize the process and influencing factors of patent examination and further discuss the methods related to patent citation recommendation and knowledge graph enhancement techniques.

Patent examination process

Patent examination is an approval process a patent must go through before it can be granted. According to the Patent Law of China, the examination and approval process of a patent application for an invention includes five stages: acceptance, preliminary examination, publication, substantive examination, and authorization. A utility model or design patent application does not undergo early publication and substantive examination in the examination and approval process. It has only three stages: acceptance, preliminary examination, and authorization (CNIPO, 2020). The examination flowchart is shown in Fig. 1.

In the patent examination process, each patent requires a minimum of two complete examinations by an examiner to evaluate the patent's novelty and so on. The typical approach to the assessment of prominent substantive features—the so-called 'three-step method’—is provided in the Guidelines for Patent Examination 2010 (CNIPO, 2010), as follows:

(i) Determine the closest prior art.
(ii) Determine the distinguishing technical features of the invention and the technical problem solved by the invention.
(iii) Determine whether the claimed invention is obvious to a person skilled in the art.

The United States Patent Office (USPTO), the European Patent Office (EPO) and the Japan Patent Office (JPO) have similar guidelines (EPO, 2023; JPO, 2015; USPTO, 2020). Patent examination is time-consuming under stringent examination requirements. In areas requiring timely technology protection, such as artificial intelligence, the average examination period is 32.81 months, which negatively impacts the protection and dissemination of new technologies (Ou et al., 2022).

The main factors influencing the length of patent examination are patent characteristics, quality, and value indicators, as well as determinants affecting the complexity of the examination task (Harhoff & Wagner, 2009). Additionally, the period of examination is also influenced by the patent agent, the number of priority claims, the length of the central claims, the number of application pages, and the examiner (Tong et al., 2018). Patent examination quality depends on examiner experience, ability, time allocated per decision, other incentives and examiner characteristics (DeGrazia et al., 2021). Increasing examiner workload can lead to systematic bias in their decisions. Examiners who need more time to conduct prior art searches are inclined to grant patents, and a large workload reduces the examination quality (Kim & Oh, 2017). However, patent search system improvements can significantly reduce appeals frequency from examiners’ rejections and grant decisions, ultimately leading to shorter examination times (Yamauchi & Nagaoka, 2015).

Therefore, improving the method of recommending patent citations at the examination stage can improve the efficiency of examination.

Patent citation recommendation

The purpose of patent citation recommendation is primarily to help patent applicants, researchers and examiners better understand and evaluate a patent’s technical background, prior art, and innovation. Depending on their target audience, they can be distinguished into three directions: patent retrieval systems for applicants (Shalaby & Zadrozny, 2019), patent analysis systems for researchers (Krestel et al., 2021), and patent recommendation systems for patent examiners (Fu et al., 2015).

For most users, access to a list of citable patents is first done through a retrieval system. And patent retrieval systems, such as Google Patents and Derwent Innovations Index, make recommendations based on the degree of similarity of patent texts. The company-oriented patent recommendation system considers the fit between the company’s needs and patent technology and recommends potentially transferable patent documents for the company by matching the company’s needs and patent technology (Chen & Deng, 2023; Lee & Sohn, 2021). The vector space model (VSM), which is based on text mining techniques, is the most commonly used, incorporating patent keyword analysis to calculate the VSM-weighted similarity of the text elements of a patent (Arts et al., 2018; Zhang et al., 2016). Incorporating semantic information in the traditional model can effectively improve the technical similarity between patents calculated by the model. Specific methods include setting different weight values to reflect the semantic information differences of words at other positions (Arts et al., 2021). Assessing the similarity among patents requires a rigorous computational analysis grounded in identifying entities within patent documents and an in-depth exploration of the semantic relations interconnecting these entities (An et al., 2021; Hain et al., 2022; Teng et al., 2024; Wang & Liu, 2022; Wang et al., 2019).

Based on the retrieval system, the patent citation recommendation system gives more consideration to the subject matter similarity or technical similarity between patents. The heterogeneous relationship between patents is modelled around subject features to extract deep patent features, and the recommended patent relevance is obviously and significantly better than the keyword-based method and the standard subject model (Chen et al., 2020). Considering multi-topic information in citation recommendations can lead to more affluent patent citation lists than similarity methods (Yin et al., 2024). It is worth mentioning that the model proposed by Choi et al., (2022a, 2022b) employs a two-stage structure, i.e., selection based on textual information and pre-trained CPC embedding values and re-ranking of candidate patents using a trained deep learning model combined with examiner citation information. The proposed model and dataset can help researchers understand the ins and outs of technical citations and better accomplish the citation recommendation task. Furthermore, in our previous study (Lu et al., 2020), we defined the concept of technical similarity between patents by categorizing patent citation relationships into patents with citations and similar but uncited patents based on the premise that patent citations are associated with knowledge. In that approach, experiments with deep learning models demonstrate that there are still technical differences between patent citations and similar patents pushed by recommender systems.

Based on the above research, we believe that it is feasible and effective to construct a citation recommendation model for patent examiners. The model can learn more in-depth information about the patent technology, enabling it to better distinguish among similar patents recommended by the retrieval system. This capability proves effective in reducing the time examiners need for patent retrieval.

Knowledge graph enhancement methods

A knowledge graph systematically portrays information comprising entities, relationships, and semantic descriptions. Entities encompass real-world objects and abstract concepts, and relationships signify the connections between entities. Semantic descriptions of entities and their relationships incorporate well-defined types and properties, each carrying a specific meaning (Ji et al., 2022). In addition to enabling the visualization of relationships between entities, knowledge graphs are used as external knowledge links in natural language processing. Knowledge graphs find applications in text augmentation by leveraging structured information to enhance and enrich textual content (Wu et al., 2022). By incorporating entities, relationships, and semantic details, knowledge graphs contribute to a more comprehensive understanding of the context, enabling improved content generation and information enrichment (Shi et al., 2023). This approach facilitates creating more contextually relevant and semantically meaningful text, enhancing textual content's overall quality and depth across various domains, from natural language processing to information retrieval, content generation systems (Dietz et al., 2018; Ridho et al., 2020).

Knowledge graphs play a significant role in the patent recommendation by leveraging structured information to enhance the relevance and efficiency of the recommendation process. By incorporating entities, relationships, and semantic details from a vast patent corpus, knowledge graphs enable a more nuanced understanding of the technological landscape. This enhanced understanding allows for identifying relevant patents based on their semantic connections, improving the accuracy of patent recommendations (Deng & Ma, 2022; Xiao et al., 2023).

Through integrating knowledge graphs, patent recommendation systems can effectively consider the intricate relationships between patents, categories, inventors, and technical concepts. This approach helps provide more precise and targeted recommendations and better interpret the recommendation results (Chen & Deng, 2023).

Therefore, in this paper, we explore the factors affecting patent examiner citation recommendation. We aim to construct a more accurate examiner citation recommendation model by developing a patent knowledge graph and incorporating the attributes and relationships of patents in the text for enhancing semantic information.

Methodology

We offer to use the knowledge graph with semantic information and the Transformer framework (Vaswani et al., 2017) with better performance to enhance Bert's[1] (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) ability to identify the patent citation relationship. We enhance the generalized K-BERT[2] (Enabling Language Representation with Knowledge Graph) (Liu et al., 2020) and introduce PK-Bert for patent examiner citation recommendations. By utilizing a patent knowledge graph, PK-BERT can seamlessly integrate domain-specific knowledge into its models, which leads to better understanding and generation of language. The overall architecture flowchart of PK-Bert is shown in Fig. 2.

Knowledge embedding layer

The knowledge embedding layer is commonly used to integrate external knowledge sources, such as knowledge graphs or ontologies, into a neural network model. This allows PK-Bert to represent words and concepts more effectively, considering their relationships and dependencies. In the context of patent analysis, the knowledge embedding layer can embed knowledge entities into the patent text and perform sentence tree transformation of the patent claim text. To begin, we define the input to our model as a patent claim text, represented as a sentence s composed of n words w_i. The $\boldsymbol{s}$ can be expressed as $s = \{w_0, w_1, w_2, \ldots, w_n\}$. Then we define a particular knowledge graph as K, which $K = \{(w_i, r_{ij}, w_{ij}), \ldots, (w_i, r_{ik}, w_{ik})\}$. (w_i, r_{ij}, w_{ij}) is a triplet, where w_i and w_{ij} are the different entities in the knowledge graph, r_{ij} is the relationship between these two entities. Finally, for the input patent sentence s, the knowledge graph K is embedded through two steps of knowledge query and knowledge injection to form the final patent sentence tree $\boldsymbol{S_{tree}}$, whose structure is shown in the Fig. 3. In the knowledge query step, if the word w_i exists corresponding to the knowledge graph entity $w_{i'}$, it is noted as (w_i, r_{i1}, w_{i1}) of w_i branch. Since a word w_i may exist in k different triples of the knowledge graph, the knowledge injection step only keeps the w_i branches of $[\big(w_i, r_{i1}, w_{i1}\big), \ldots, (w_i, r_{ik}, w_{ik})]$ but does not

[1] https://github.com/google-research/bert.

[2] https://github.com/autoliuweijie/K-BERT.

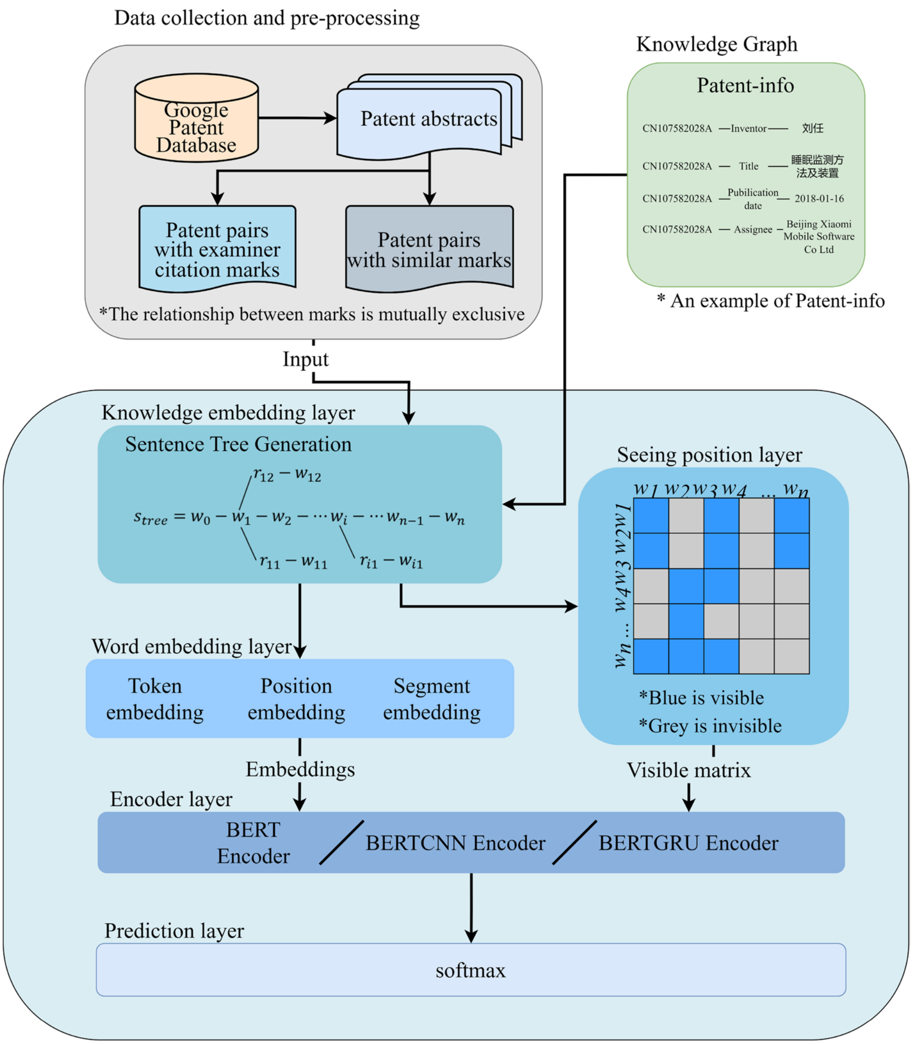
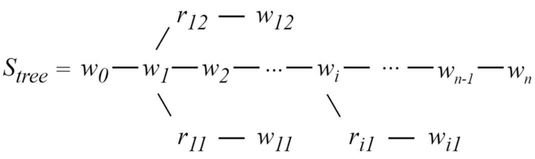

Data collection and pre-processing
Google Patent Database
Patent abstracts
Patent pairs with examiner citation marks
Patent pairs with similar marks
*The relationship between marks is mutually exclusive
Knowledge Graph
Patent-info
CN107582028A —Inventor— 刘任
CN107582028A — Title — 睡眠监测方法及装置
CN107582028A — Publication date — 2018-01-16
CN107582028A — Assignee — Beijing Xiaomi Mobile Software Co Ltd
* An example of Patent-info
Input
Knowledge embedding layer
Sentence Tree Generation
$s_{tree} = w_0 - w_1 - w_2 - \cdots w_i - \cdots w_{n-1} - w_n$
$r_{12} - w_{12}$
$r_{11} - w_{11}$
$r_{i1} - w_{i1}$
Seeing position layer
$w_1\ w_2\ w_3\ w_4\ \ldots\ w_n$
*Blue is visible
*Grey is invisible
Word embedding layer
Token embedding
Position embedding
Segment embedding
Embeddings
Visible matrix
Encoder layer
BERT Encoder
BERTCNN Encoder
BERTGRU Encoder
Prediction layer
softmax

Fig. 2 Data preparation and model structure

Fig. 3 The structure of the sentence tree S_{tree}


$S_{tree} = w_0 — w_1 — w_2 — \cdots — w_i — \cdots — w_{n\text{-}1} — w_n$
$r_{12} — w_{12}$
$r_{11} — w_{11}$
$r_{i1} — w_{i1}$

consider continuing to operate knowledge query of w_{ik}. That is each word w_i can have more than one tree branch, but the depth of the tree branch is at most one.

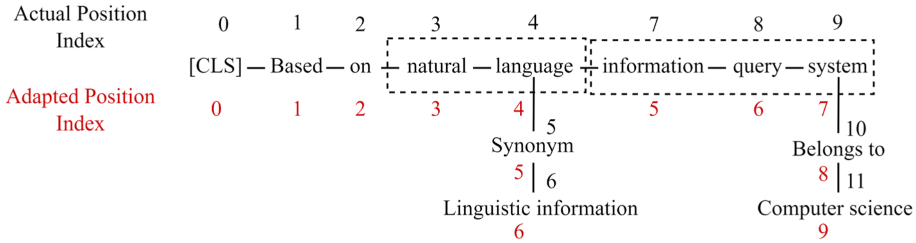


Fig. 4 The positional index of each node in the sample sentence tree is used to guide the injection of knowledge entities

Seeing position layer

Since the addition of knowledge graph entities may change both the word order and semantic logic of the original sentence, addressing the knowledge noise problem is also a very critical issue. Due to the embedding of knowledge, it is also essential to solve the problem of spatial consistency between the knowledge vector and the original text vector. This study here refers to the seeing layer in the generic K-Bert model. It operates the visual position representation of patent sentence trees to overcome the problem of semantic logic confusion brought by incorporating knowledge graphs.

Let's take the "Information query system based on natural language" for example. The sentence tree structure is shown in Fig. 4. The actual position index information is marked in black. But the position information of the branch should be closely associated with the corresponding word, so the adapted position index information is kept in red.

We build the corresponding seeing position matrix according to the sample sentence tree. From the black position numbers in Fig. 4. The "Synonym" at position 5 and the "Linguistic information" at position 6 are only associated with the "natural" and the "language" at position 3 and 4. Therefore, positions 5 and 6 are visible with positions 3, 4, 5, 6 (marked as blue), and invisible with positions 0, 1, 2, 7, 8, 9, 10, 11 (marked as grey). Following this logic, we eventually obtain the 0–1 visual position matrix as shown in Fig. 5.

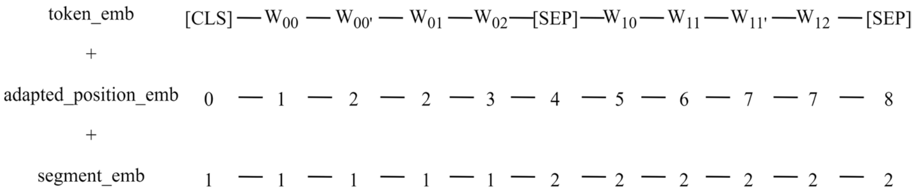
Word embedding layer

The role of the word embedding layer is to transform the output sentence tree of the knowledge layer into the word embeddings. Referring to the embedding representation of the Bert model, the word embedding here also consists of token embedding, adapted-position embedding and segment embedding. Since the sentence structure is altered after incorporating knowledge, preserving the original structural information while transforming the modified sentence tree into a word sequence is the crucial point of this layer. Taking two patent texts $A: \{w_{00}, w_{00\prime}, w_{01}, w_{02}\}$ and $B: \{w_{10}, w_{11}, w_{11\prime}, w_{12}\}$ as the input examples of the patent examiner citation recommendation model, the final obtained patent text input vector is $X = token\_emb + adapted\_position\_emb + segment\_emb$, as shown in Fig. 6.

For token embedding, to obtain a suitable word sequence, word embedding layer reorders the words in the sentence tree, i.e., the supplementary knowledge words of the branches are inserted after the corresponding original words, and the following words are sequentially arranged backward in order. In addition, some flag bits with notable

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	1	1	1	1	0	0	1	1	1	0	0
1	1	1	1	1	1	0	0	1	1	1	0	0
2	1	1	1	1	1	0	0	1	1	1	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0
5	0	0	0	1	1	1	1	0	0	0	0	0
6	0	0	0	1	1	1	1	0	0	0	0	0
7	1	1	1	1	1	0	0	1	1	1	1	1
8	1	1	1	1	1	0	0	1	1	1	1	1
9	1	1	1	1	1	0	0	1	1	1	1	1
10	0	0	0	0	0	0	0	1	1	1	1	1
11	0	0	0	0	0	0	0	1	1	1	1	1

Fig. 5 Seeing position matrix for sample sentence tree



token_emb [CLS] — W_{00} — $W_{00'}$ — W_{01} — W_{02} — [SEP] — W_{10} — W_{11} — $W_{11'}$ — W_{12} — [SEP]

+

adapted_position_emb 0 — 1 — 2 — 2 — 3 — 4 — 5 — 6 — 7 — 7 — 8

+

segment_emb 1 — 1 — 1 — 1 — 1 — 2 — 2 — 2 — 2 — 2 — 2

Fig. 6 Word embedding of patent text input sample

roles are added to the input of the Bert model. Among them, the [CLS] flag is usually placed at the first place of a sentence, indicating that the corresponding sentence representation vector is a classification task. And the [SEP] flag is used to separate the two sentences of the input. For the above input two patent texts A and B, the token embedding of this paper is $\{$[CLS], $w_{00}, w_{00\prime}, w_{01}, w_{02}$, [SEP], $w_{10}, w_{11}, w_{11\prime}, w_{12}$, [SEP]$\}$.

For adapted-position embedding, from the seeing mentioned above position layer, it is known that the sentence tree structure incorporated into the knowledge graph has two types of the actual and adapted position index. Here we choose the adjusted position index. That's because the entity of the knowledge embedding is only related to the associated words, so the position of the knowledge embedding should be after the related terms. Taking the above input example, for patent A, $w_{00'}$ is the knowledge embedding word of w_{00}, while w_{01} is the next word of w_{00} according to the original sentence order, so the adapted position index of $w_{00'}$ and w_{01} are the same.

For segment embedding, it is used to mark the different sentences when input contains multiple sentences. When two input sentences are used for the semantic matching task, the segment embedding of the above input example is {1, 1, …, 1, 2, 2, …, 2}.

Encoder layer

The encoder layer mainly considers combining the Bert model with better encoding performance and general neural network models such as convolutional neural networks (CNN) and gate recurrent unit (GRU) to obtain the corresponding Bert, BertCNN and BertGRU encoder by splicing between these models. In this section, we use universal encoder representations (UER)[3] (Zhao et al., 2019) to implement splicing between different neural networks.

Bert encoder mainly consists of a multi-headed self-attention mechanism, a fully connected feedforward neural network layer, and residual connectivity and normalization. First, we transform the patent text incorporated into the knowledge graph into the vector–matrix through word embedding. Then, in the self-attention layer, to express the semantics at a deeper level, different linear variations of the text vector–matrix $\boldsymbol{X}$ are performed to obtain the corresponding $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. The calculation formula is shown in Eqs. (1)–(3).

$$Q = Linear(X) = X * W^{Q} \tag{1}$$

$$K = Linear(X) = X * W^{K} \tag{2}$$

$$V = Linear(X) = X * W^{V} \tag{3}$$

Further, to learn different aspects of text features, PK-Bert considers the original multi-headed attention mechanism output and the seeing position layer influence, so the final attention mechanism formula is shown in Eq. (4) below. The $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the matrices obtained from the above linear variation, $\boldsymbol{M}$ is the seeing position matrix above, and d_k is the scale factor used to offset the effect of the dot product calculation.

$$Attention = softmax\left(\frac{QK^{T} + M}{\sqrt{d_k}}\right) * V \tag{4}$$

Since the format of the vector after the Bert encoder is "[CLS]{Patent A text vector}[SEP]{Patent B text vector}[SEP]", we select the [CLS] corresponding vector of the last layer as the output vector of the patent text pair.

As for BertCNN, after the Bert encoder, the CNN encoder conducts convolution and pooling operations on the $n \times k$-dimensional text vector matrix. Different convolution kernel sizes $h_i \times k \times m$, where h_1, h_2, and h_3 are the heights of three different convolution kernels, k is their width, which is consistent with the dimensionality of the word embedding,

[3] https://github.com/dbiir/UER-py.

and *m* is the number of each kernel, are employed to extract semantic features from various aspects of the text. The convolution operation divides the matrix into windows, and for each window, the ReLU activation function is applied to obtain feature vectors. The resulting feature vectors are further processed through pooling operations, selecting the maximum value for each feature to form a pooling vector output. This reduces model complexity and transforms the convolutional layer output into a fixed-length input for the classification layer. Finally, the maximum pooling vectors are vertically stitched to create the final splicing vector, denoted as c_{max}, the output vector characterizing the patent text pair.

Like the BertCNN encoder, the $n \times k$-dimensiona vector matrix V for the patent text pair is obtained through the Bert encoder. We utilize the bidirectional GRU model for context encoding to optimize computational efficiency while maintaining effectiveness. The update and reset gate input vectors filter and update the hidden state based on the input sequence X_t at moment t. The input information for updating the hidden state is obtained through the tanh function applied to the filtered hidden state h_{t-1}. The final hidden state $h_t\prime$ at time t is obtained through weighted summation. The bidirectional GRU output is a vertically spliced combination of the forward $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ coding outputs, denoted as h_t. Given that the last bit of bidirectional GRU encoding encapsulates all semantic information, it is selected as the output vector characterizing the patent text pair.

Prediction layer

After Encoder encoding, the final splicing vector of the full-connected layer stitching is obtained. We then output the probability values of the two classes by the Softmax classifier and take the class where the maximum probability value is located as the predicted classification class.

The fully connected layer mainly maps the features learned in the Encoder layer to a one-dimensional equal-length vector space to facilitate subsequent classification operations. The output vector of the fully connected layer is then vector normalized by the Softmax layer, and the calculation equations are shown in Eqs. (5) and (6).

$$z_i = w_i * x + b_i \tag{5}$$

$$y_i = \frac{e^{z_i}}{\sum_{i=1}^{3} e^{z_i}} \tag{6}$$

Experiment

This section presents experiments that evaluate the validity of PK-Bert against each other. Specifically, we aim to answer the following evaluation questions:

Question 1: Does fusing knowledge graphs improve the performance of patent examiner citation recommendations?

Question 2: Which part of the patent knowledge graph is more helpful for examiner citation recommendation?

Question 3: Can models incorporating knowledge graphs achieve better results using earlier training data on the latest test set?

Table 1 Experimental dataset situation

Label	GPair-NLP	GPair-H04
Examiner cited patent pairs (1)	19,811	22,149
Uncited but similar patent pairs (0)	19,811	22,149

Data collection and knowledge graph construction

Data collection

We obtained patent data in two fields in Chinese from Google Patents. Patents are classified, and each classification number is examined to ensure accuracy. When searching for a patent, one can use either keywords or classification numbers. We collected data using keywords and classification numbers to construct a dataset for model validation. This approach covers many search cases and yields data with distinct characteristics. Furthermore, considering the input data length constraints of the Bert model, we opted for shorter patent abstracts as input text.

In our dataset, each patent is linked to relevant patents through citations, being cited by others, and thematic similarity. The data collection involves initial keyword searches on the Google Patents website to compile comprehensive patent lists. Subsequently, we systematically access the details page for each patent, where Google Patents provides three lists: citations, cited by, and similar patents. The similarity list is generated by the Google Patents system, and in both the citation and cited by lists, examiner-provided citations are distinguished. Through meticulous data cleaning, we identified no overlap between patents in the citation and similar lists. We curated patents meeting our criteria from these lists, forming pairs as follows: Patent A-Patent B-Label 0 (uncited but similar), Patent A-Patent C-Label 1 (examiner citation).

The first dataset comprises a collection of 10,030 patent abstract texts acquired using the keyword 'natural language processing'. Meanwhile, the second dataset consists of 17,586 patent abstract texts obtained by filtering patents with the IPC number set to 'H04'. The final experimental dataset of this paper includes Google Patent Pair of Natural Language Processing (GPair-NLP) and Google Patent Pair of H04 classification (GPair-H04).

And we adopt a quantitative screening approach to guarantee both balance and completeness in the dataset. During the quantitative screening phase, illustrated with GPair-NLP, we initially amassed around 10,000 patents, yielding approximately 50,000 pairs of patent relationships. Following additional statistical analysis and data cleaning, we excluded pairs with a single relationship type. Post-cleaning, we noted a slight surplus of similar patent pairs over examiner-cited pairs. Consequently, in the second round of data collection, we concentrated solely on patents where the number of examiner-cited pairs surpassed that of similar pairs, continuing until a balanced quantity of patents for both relationship types was achieved.

The experimental training set, validation set: test set = 8:1:1 is set. The basic statistics are shown in Table 1. And the distribution of patent publication dates within the datasets are illustrated in Fig. 7.

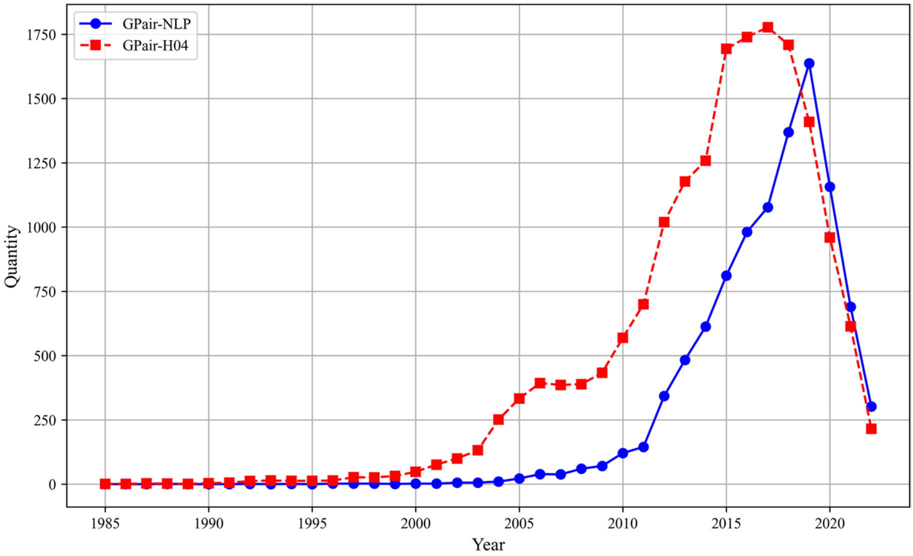


Fig. 7 Distribution of patent publication time in the dataset

Patent knowledge graph construction

In the knowledge graph embedding part, we choose CnDBpedia[4] (Xu et al., 2017) as knowledge graphs for generic domains. CnDBpedia is a knowledge graph based on the DBpedia[5] ontology and covers a wide range of domains, including geography, history, culture, and science. It contains over 4.5 million entities and 5.1 million resource description framework (RDF) triples, making it one of the largest knowledge graphs in the world. CnDBpedia is constructed by extracting structured information from Chinese Wikipedia, and it is continuously updated to reflect changes in the underlying Wikipedia pages.

In order to validate the advantages and disadvantages of generic and domain knowledge graphs for the semantic representation of patents, we focus on constructing a comprehensive knowledge graph centered around patent information. Employing the patent number as a unique identifier, we gathered five essential attributes—patent title, IPC classification, inventors, assignee, and publication date—to form the foundation of our knowledge graph. During the knowledge graph embedding phase, to facilitate seamless embedding, we prepended each original patent text with its corresponding patent number. The constructed Patent-info knowledge graph adheres to a (Patent Number, Attribute, Attribute Value) format, ensuring that by using the patent number as an index, the corresponding patent information can be effectively embedded. The number of entities of each type is shown in Table 2.

[4] http://kw.fudan.edu.cn/cndbpedia/intro/.

[5] https://dbpedia.org/page/DBpedia.

Table 2 The number of entities of each type

Knowledge graphs	#Entities	#Relations	#Triples
CnDBpedia	4,597,165	62,278	5,168,865
Patent-info	103,973	5	215,089

Experimental model settings

Experimental model introduction

Our experiments include the following models.

(i) Text CNN/GRU (Lu et al., 2020): The Siamese twin network structure is used to first encode the same CNN/GRU for each of the two patents, and then output an equal-length structural vector after the dot product and dot subtraction computation methods. Finally, the resulting text representation vector is input to Softmax to predict the result of the patent pair.

(ii) Bert/BertCNN/BertGRU: The Bert model based on Transformer encoding is proposed for the patent text semantic matching task. The special flags such as [CLS] and [SEP] in the Bert model are used to process the patent pair text, and the encoding vector of the patent pair text is obtained and finally input to the Softmax classifier to predict the result of the patent pair.

(iii) K-Bert/K-BertCNN/K-BertGRU: Knowledge embedding employs the Bert model with CnDBpedia.

(iv) PK-Bert/PK-BertCNN/PK-BertGRU: Knowledge embedding employs the Bert model with Patent-info.

Experimental setup and model training

The neural network model usually contains many parameters, among which the model's internal parameters, such as weight values, are updated automatically during training. In contrast, the model's initialization parameters must be set by oneself before training. Here, through pre-experiments and reference to previous experiences, the main parameters, parameter descriptions, and settings involved in the model experiments are shown in Appendix A.

Patent examiner citation recommendation

Models evaluation (question 1)

Considering that the number of samples of each category in the patent dataset in this study is not the same, to overcome the influence of different category shares on measuring model effects, the macro-average metrics of precision, recall, and F1 value are used as evaluation metrics. The experimental results of all models in the patent datasets of the two datasets are shown in Table 3. We mark the best experimentalresults using bold.

Table 3 Experimental results on the GPair-NLP and GPair-H04

Models	GPair-NLP			GPair-H04		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Text-CNN	73.30	73.20	73.30	78.20	78.00	78.00
Text-GRU	74.30	74.25	74.2%	78.25	78.15	78.15
Bert	87.10	87.10	87.10	86.70	86.60	86.60
BertCNN	88.00	87.80	87.90	87.50	87.45	87.45
BertGRU	88.20	88.20	88.25	87.25	87.15	87.20
K-Bert (with CnDBpedia)	87.60	87.60	87.60	86.90	86.85	86.90
PK-Bert (with Patent-info)	88.40	88.40	88.40	87.75	87.70	87.75
K-BertCNN (with CnDBpedia)	88.10	87.90	88.00	87.55	87.25	87.30
PK-BertCNN (with Patent-info)	89.10	88.90	89.00	87.90	87.90	87.90
K-BertGRU (with CnDBpedia)	88.10	88.00	88.00	87.50	87.45	87.45
PK-BertGRU (with Patent-info)	**89.25**	**89.25**	**89.25**	**87.95**	**88.00**	**88.00**

The K-Bert series models consistently outperform the baseline Bert model across all evaluated metrics (precision, recall, and F1 scores). This indicates that integrating knowledge graphs within the Bert series enhances examiner citation recommendation task performance.

Notably, the Patent-info knowledge graph contains a modest entity count of 103,793, significantly fewer than the 4,597,165 entities present in CnDBpedia; the former's quantity is merely 2% of the latter. However, models incorporating the Patent-info knowledge graph demonstrate superior performance to those using CnDBpedia. This trend emphasizes the importance of leveraging patent-specific information for practical examiner citation recommendations within the patent domain.

The patent data often includes sequences of technical terms and concepts that benefit from the sequential processing capabilities of the GRU architecture. GRU excels in capturing the nuanced relationships and dependencies between words and phrases within the patent text, adequately representing the sequential nature inherent in patent information. Conversely, CNN is better suited for capturing hierarchical features and patterns but might need to be more effective in handling the sequential dependencies prevalent in patent texts.

In the context of examiner citation recommendation, where the chronological order of information is crucial, the GRU architecture is advantageous. It excels in understanding the context of prior citations and the sequence of information in patents, making it well-suited for this task.

To further analyze the experimental results, we present the experimental results of the two datasets output by category in Tables 4 and 5. We mark the best experimental results using bold.

Models exhibit consistently higher recall and F1 scores for Label 1 across both datasets. This trend can be attributed to the nature of the task, where Label 1 likely represents positive instances or relevant citations. The emphasis on correctly identifying and recommending relevant citations aligns with the task's objective, leading to higher recall and F1 scores for Label 1. The model's capacity to effectively capture and recommend relevant citations contributes to this superior performance.

Table 4 Experimental results by category on the GPair-NLP

Models	GPair-NLP					
	Uncited but similar patent pairs (0)			Examiner cited patent pairs (1)		
	Precision (%)	Recall (%)	F1(%)	Precision (%)	Recall (%)	F1 (%)
Text-CNN	73.00	74.50	73.70	73.30	71.80	72.50
Text-GRU	75.10	72.40	73.70	73.50	76.10	74.70
Bert	88.10	86.10	87.10	86.10	88.10	87.10
BertCNN	90.00	85.00	87.42	85.90	90.60	88.20
BertGRU	88.90	87.60	88.30	87.50	88.80	88.20
K-Bert (with CnDBpedia)	89.00	85.90	87.50	86.10	89.20	87.60
PK-Bert (with Patent-info)	89.30	87.40	88.30	87.40	89.30	88.30
K-BertCNN (with CnDBpedia)	90.10	85.20	87.60	86.00	90.60	88.30
PK-BertCNN (with Patent-info)	**91.80**	85.40	88.50	86.40	**92.40**	**89.30**
K-BertGRU (with CnDBpedia)	90.00	85.50	87.70	86.20	90.50	88.30
PK-BertGRU (with Patent-info)	89.20	**89.20**	**89.20**	**89.30**	89.30	89.30

Table 5 Experimental results by category on the GPair-H04

Models	GPair-H04					
	Uncited but similar patent pairs (0)			Examiner cited patent pairs (1)		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Text-CNN	79.80	74.60	77.10	76.60	81.40	78.90
Text-GRU	79.70	75.60	77.60	76.80	80.70	78.70
Bert	87.80	84.80	86.20	85.60	88.40	87.00
BertCNN	86.90	88.00	87.40	88.10	86.90	87.50
BertGRU	88.20	85.60	86.90	86.30	88.70	87.50
K-Bert (with CnDBpedia)	87.60	85.60	86.60	86.20	88.10	87.20
PK-Bert (with Patent-info)	88.40	86.50	87.50	87.10	88.90	88.00
K-BertCNN (with CnDBpedia)	**90.00**	83.70	86.70	85.10	**90.80**	87.90
PK-BertCNN (with Patent-info)	87.00	**88.80**	**87.90**	**88.80**	87.00	87.90
K-BertGRU (with CnDBpedia)	88.50	85.80	87.10	86.50	89.10	87.80
PK-BertGRU (with Patent-info)	88.50	87.30	**87.90**	87.40	88.70	**88.10**

Despite the higher scores for Label 1, the evaluation metrics for Label 0 do not exhibit imbalance, ensuring a reliable overall assessment of the model's performance. The absence of significant disparities in precision, recall, and F1 scores between the two labels indicates that the model maintains a balanced approach in handling positive and negative instances. This balance is crucial for ensuring that the model's recommendations are not skewed towards a particular label, thus enhancing the reliability of the overall evaluation results.

Table 6 Experimental results on the GPair-NLP and GPair-H04

Models	GPair-NLP		GPair-H04	
	Recall	F1	Recall	F1
Patent-info	88.40%	88.40%	87.70%	87.75%
Patent-info without assignee	88.05% (– 0.35%)	88.00% (– 0.40%)	87.30% (– 0.40%)	87.35% (– 0.40%)
Patent-info without classification	88.25% (– 0.15%)	88.25% (– 0.15%)	87.65% (– 0.05%)	87.70% (– 0.05%)
Patent-info without publication date	88.65% (+0.25%)	88.60% (+0.20%)	88.20% (+0.50%)	88.15% (+0.40%)
Patent-info without inventor	87.90% (-0.50%)	87.90% (– 0.50%)	86.85% (– 0.85%)	86.85% (– 0.90%)
Patent-info without title	88.70% (+0.30%)	88.70% (+0.30%)	87.80% (+0.10%)	87.80% (+0.05%)

Table 7 Latest test dataset details

Label	Count
Examiner cited patent pairs (1)	2718
Uncited but similar patent pairs (0)	4290

Ablation study (question 2)

To gain deeper insights into the factors that exert the most significant influence on the task of examiner citation recommendation within the patent knowledge graph, we conducted ablation experiments using the K-Bert model as our foundational framework. The experimental results are presented in Table 6, offering a comprehensive view of the effects of various factors on the performance of the citation recommendation task.

Removing the publication date and title improves model performance, indicating these elements might not be crucial and could even introduce noise. In contrast, excluding inventor details has a substantial impact, underlining its pivotal role in the model's recommendation capabilities. In addition, assignees also has some impact on the results of the experiment. Patent categorization, although less influential, still plays a role.

Patents from the same inventor tend to concentrate on specific technological features, rendering them highly likely to exhibit technical similarities. For patent examiners, this signifies that patents from a joint inventor are particularly relevant due to their heightened probability of sharing similar technologies. Moreover, it is common for inventors to utilize technologies akin to their previously published patents, strategically refraining from citing their work to enhance the likelihood of patent authorization. Additionally, assignees, often corporate entities, tend to focus on acquiring patents within a specific domain. Consequently, patents held by the same assignee are more likely to cover a wide range of product technologies. Lastly, patents classified under the same IPC code generally share similar technological aspects, increasing the likelihood of citation by examiners.

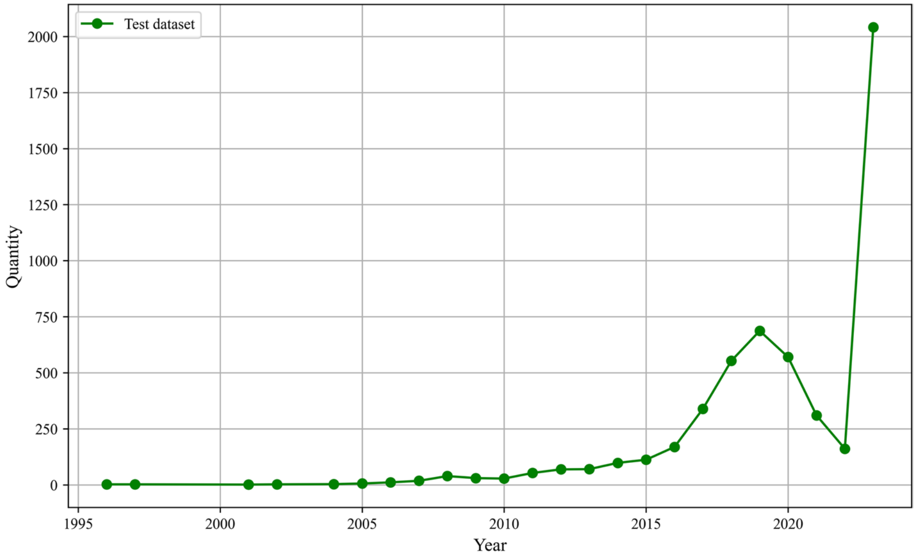


Fig. 8 Distribution of publication time in the latest test set

Table 8 Latest patent dataset experimental results

Models	Latest patent dataset		
	Precision (%)	Recall (%)	F1 (%)
Text-CNN	53.05	52.40	51.55
Text-GRU	58.60	55.10	55.74
Bert	70.25	69.35	69.70
BertCNN	70.90	68.60	69.15
BertGRU	71.10	67.95	68.50
K-Bert (with CnDBpedia)	71.15	69.50	69.95
PK-Bert (with Patent-info)	71.95	70.70	71.10
K-BertCNN (with CnDBpedia)	71.10	69.60	70.05
PK-BertCNN (with Patent-info)	71.70	71.10	71.30
K-BertGRU (with CnDBpedia)	71.85	70.55	71.00
PK-BertGRU (with Patent-info)	**72.20**	**71.25**	**71.60**

Assessing models on latest patents (question 3)

This part of our study investigates if models, equipped with knowledge graphs, can perform better when trained on earlier data and tested on the latest patents from November 2023 in artificial intelligence. We randomly collected 1,869 patents with publication dates after November 2023 and captured their examiner citations and lists of similar patents to form the test set. The test set details are shown in Table 7, and the distribution of patent issuance times involved in the dataset is shown in Fig. 8.

As can be seen in Fig. 8, the test patents are concentrated in 2023, while the latest datasets used for model training are published in 2022, covering the years from 2014 to 2018. There is no intersection between the training and test sets. The experimental results are shown in Table 8. We mark the best experimental results using bold.

Based on the results of the experiments conducted on the latest patent dataset, we can draw several observations regarding recall values:

(i) Text-CNN and Text-GRU models show lower recall values (52.40% and 55.10%, respectively). The convolutional and recurrent structures may not capture the intricate relationships within the patent text as effectively as knowledge-enhanced models.

(ii) Bert, BertCNN, and BertGRU models demonstrate improved recall values, with BertGRU being slightly lower. Bert-based models leverage contextual information effectively but may face challenges in capturing nuanced patterns within patent texts.

(iii) PK-Bert models consistently outperform other models in recall. This indicates that incorporating patent-specific knowledge enhances the models' ability to identify relevant citations, resulting in higher recall. It underscores the importance of utilizing domain-specific knowledge graphs to improve citation recommendations.

Conclusion and future works

Our research utilizes knowledge graphs to augment examiners' citation recommendations. Notably, despite the small number of entities in the Patent-info knowledge graph (only 2% of CnDBpedia), it consistently outperforms models using the CnDBpedia, highlighting the importance of using patent-specific information for practical recommendations and emphasizing the importance of domain-specific knowledge graphs. The Experiments on latest patents show that knowledge-enhanced models, especially those containing Patent-info, have a sustained advantage in terms of recall. It affirms the efficacy of domain-specific knowledge graphs and provides valuable insights for patent examiners seeking more granular and context-aware citation recommendations.

However, it is essential to recognize some limitations of our study. First, the depth of our analysis of this category is limited by the need for a publicly available dataset of Chinese patent examiner citations. This limitation prevents us from gaining a more comprehensive understanding of the results of Chinese patent recommendations and their specific categories.

Then, our study does not cover patent examiner citations from different countries or the lack of analysis of patents of the same family. In addition, while showing value, the Patent-info Knowledge Graph has limitations. It mainly consists of structured attributes and lacks deeper semantic information.

Lastly, after incorporating the knowledge graph, our recommendation model can provide a certain degree of interpretability. However, our results are still general for patent examination, which requires a detailed reasoning process.

Potential enhancements to the knowledge graph involve integrating additional semantic features to further augment the model's comprehension of patent-related information. The utilization of graph neural networks for embedding is under consideration in the knowledge embedding component (Choi et al., 2022a, 2022b; Choi & Yoon, 2022). Additionally, taking into account the features cited by examiners from different national patent offices will contribute to enhancing the interpretability of the model.

Appendix

See Table 9.

Table 9 Description and settings of the main parameters

Parameters name	Parameters description	Parameters setting
Max_seq_length	Maximum sentence length	512
Kg_name	Embedded knowledge graph name or path	CnDBpedia, Patent-info
Max_entities	Maximum number of connected entities	2
Visible_matrix	Whether to use visual location matrix	True
embedding_size	Dimensionality of word embedding	768
Heads_num	Number of heads of attention mechanisms in the Bert encoding module	12
Layers_num	Number of layers in the Bert encoding module	12
Hidden_size	Dimension of hidden layers in the Bert encoding module	768
Rnn_hidden	Dimensionality of the hidden state vector in GRU coding	768
Kernel_size	Shape of convolutional kernel in CNN coding	[2, 3, 4]
Learning_rate	Learning rate	2e − 5
Warm_up	Warm-up learning rate	0.1
Batch_size	Number of training/validation/test samples per batch	32
Epochs_num	Number of iterations	10
Optimizer	Optimization algorithm	Adam
Loss	Loss function	CrossEntropy
Dropout	Dropout settings	0.5

Acknowledgements The authors warmly thank reviewers for their valuable suggestions. This research was partly supported by Key-Area Research and Development Program of Guangdong Province (NO.2021B0101420004), National Natural Science Foundation of China (NO.71801229).

Author contribution YL: Investigation, Writing—review and editing. XT: Methodology, Investigation, Writing—review and editing. XX: Methodology, Software, Writing—original draft. HZ: Supervision, Funding acquisition, Writing—review and editing.

Funding Funding was provided by Key-Area Research and Development Program of Guangdong Province (Grant No. 2021B0101420004), National Natural Science Foundation of China (Grant No. 71801229).

Declarations

Competing interests The authors declare no competing interests.

References

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics, 88*(4), 774–779.

An, X., Li, J., Xu, S., Chen, L., & Sun, W. (2021). An improved patent similarity measurem-ent based on entities and semantic relations. *Journal of Informetrics, 15*(2), 101135. https://doi.org/10.1016/j.joi.2021.101135

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal, 39*(1), 62–84. https://doi.org/10.1002/smj.2699

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy, 50*(2), 104144. https://doi.org/10.1016/j.respol.2020.104144

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223–229. https://doi.org/10.1002/asi.4630360402

Chen, H., & Deng, W. (2023). Interpretable patent recommendation with knowledge graph and deep learning. *Scientific Reports, 13*(1), 2586. https://doi.org/10.1038/s41598-023-28766-y

Chen, J., Chen, J., Zhao, S., Zhang, Y., & Tang, J. (2020). Exploiting word embedding for he-erogeneous topic model towards patent recommendation. *Scientometrics, 125*(3), 2091–2108. https://doi.org/10.1007/s11192-020-03666-4

Chen, L. (2017). Do patent citations indicate knowledge linkage? The evidence from text simil-arities between patents and their citations. *Journal of Informetrics, 11*(1), 63–79. https://doi.org/10.1016/j.joi.2016.04.018

China National Intellectual Property Office (CNIPO). (2010). *Guidelines for patent examination 2010*. Intellectual Property Publishing House Co., Ltd.

China National Intellectual Property Office (CNIPO). (2020). *Process of patent application examination and approval*. Retrieved December 10, 2023, from https://www.cnipa.gov.cn/art/2020/6/5/art_1517_92471.html

Choi, J., & Yoon, J. (2022). Measuring knowledge exploration distance at the patent level: Application of network embedding and citation analysis. *Journal of Informetrics, 16*(2), 101286. https://doi.org/10.1016/j.joi.2022.101286

Choi, J., Lee, J., Yoon, J., Jang, S., Kim, J., & Choi, S. (2022a). A two-stage deep learning- based system for patent citation recommendation. *Scientometrics, 127*(11), 6615–6636. https://doi.org/10.1007/s11192-022-04301-0

Choi, S., Lee, H., Park, E., & Choi, S. (2022b). Deep learning for patent landscaping using transformer and graph embedding. *Technological Forecasting and Social Change, 175*, 121413. https://doi.org/10.1016/j.techfore.2021.121413

DeGrazia, C. A. W., Pairolero, N. A., & Teodorescu, M. H. M. (2021). Examination incentives, learning, and patent office outcomes: The use of examiner's amendments at the USPTO. *Research Policy, 50*(10), 104360. https://doi.org/10.1016/j.respol.2021.104360

Deng, W., & Ma, J. (2022). A knowledge graph approach for recommending patents to companies. *Electronic Commerce Research, 22*(4), 1435–1466. https://doi.org/10.1007/s10660-021-09471-2

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Dietz, L., Kotov, A., & Meij, E. (2018). Utilizing knowledge graphs for text-centric information retrieval. In: *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1387–1390). https://doi.org/10.1145/3209978.3210187

European Patent Office (EPO). (2023). *Guidelines for examination*. Retrieved December 10, 2023, from https://www.epo.org/en/legal/guidelines-epc/2023/index.html

Färber, M., & Jatowt, A. (2020). Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries, 21*, 375–405. https://doi.org/10.1007/s00799-020-00288-2

Fu, T.Y., Lei, Z., & Lee, W.C. (2015). Patent Citation Recommendation for Examiners. In *Proceedings of the 2015 IEEE international conference on data mining (ICDM)* (pp. 751–756). https://doi.org/10.1109/ICDM.2015.151

Hain, D. S., Jurowetzki, R., Buchmann, T., & Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change, 177*, 121559. https://doi.org/10.1016/j.techfore.2022.121559

Harhoff, D., & Wagner, S. (2009). The duration of patent examination at the European patent office. *Management Science, 55*(12), 1969–1984. https://doi.org/10.1287/mnsc.1090.1069

Japan Patent Office (JPO). (2015). *Examination guidelines for patent and utility model in Japan.* Retrieved December 10, 2023, from https://www.jpo.go.jp/e/system/laws/rule/guideline/patent/tukujitu_kijun/index.html

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems, 33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Kim, Y. K., & Oh, J. B. (2017). Examination workloads, grant decision bias and examination quality of patent office. *Research Policy, 46*(5), 1005–1019. https://doi.org/10.1016/j.respol.2017.03.007

Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information, 65*, 102035. https://doi.org/10.1016/j.wpi.2021.102035

Kuhn, J. M. (2011). Information overload at the U.S. patent and trademark office: Reframing the duty of disclosure in patent law as a search and filter problem. *Journal of Law and Technology, 13*(3), 90–139. https://doi.org/10.1016/j.wpi.2021.102035

Lee, J., & Sohn, S. Y. (2021). Recommendation system for technology convergence opportunit-ies based on self-supervised representation learning. *Scientometrics, 126*(1), 1–25. https://doi.org/10.1007/s11192-020-03731-y

Lin, D., Sun, J., Hao, T., & Wang, C. (2016). Research on the applicability of patent citation in patent value evaluation. *Journal of Intelligence, 35*(12), 150–154.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020). K-BERT: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(03), 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681

Lu, Y., Xiong, X., Zhang, W., Liu, J., & Zhao, R. (2020). Research on classification and simi-larity of patent citation based on deep learning. *Scientometrics, 123*, 813–839. https://doi.org/10.1007/s11192-020-03385-w

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics, 49*(1), 93–123. https://doi.org/10.1023/A:1005613325648

Ou, G., Pang, N., & Wu, J. (2022). Influencing factors of patent examination cycle: Case study of artificial intelligence in China. *Data Analysis and Knowledge Discovery, 8*(6), 20–30.

Ridho, R., Edgar, M., & Maarten, D. R. (2020). *Knowledge graphs: An information retrieval perspective*. Now Foundations and Trends.

Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: A literature review. *Knowledge and Information Systems, 61*(2), 631–660. https://doi.org/10.1007/s10115-018-1322-7

Shi, K., Cai, X., Yang, L., & Zhao, J. (2023). Enriched entity representation of knowledge graph for text generation. *Complex & Intelligent Systems, 9*(2), 2019–2030. https://doi.org/10.1007/s40747-022-00898-0

Teng, H., Wang, N., Zhao, H., Hu, Y., & Jin, H. (2024). Enhancing semantic text similarity with functional semantic knowledge (FOP) in patents. *Journal of Informetrics, 18*(1), 101467. https://doi.org/10.1016/j.joi.2023.101467

Tong, T. W., Zhang, K., He, Z., & Zhang, Y. (2018). What determines the duration of patent examination in China? An outcome-specific duration analysis of invention patent applications at SIPO. *Research Policy, 47*(3), 583–591. https://doi.org/10.1016/j.respol.2018.01.002

United States Patent Office (USPTO). (2020). *Understanding the patent examination process.* Retrieved December 10, 2023, from https://www.uspto.gov/sites/default/files/documents/InventionCon2020_Understanding_the_Patent_Examination_Process.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Curran Associates Inc

Wada, T. (2018). The choice of examiner patent citations for refusals: Evidence from the trilateral offices. *Scientometrics, 117*(2), 825–843. https://doi.org/10.1007/s11192-018-2885-5

Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y., & Huang, Y. (2019). Measuring patent similarity with SAO semantic analysis. *Scientometrics, 121*(1), 1–23. https://doi.org/10.1007/s11192-019-03191-z

Wang, Z., & Liu, Y. (2022). SEA-PS: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*. https://doi.org/10.1177/01655515221106651

Wu, J., Li, B., Ji, Y., Tian, J., & Xiang, Y. (2022). Text-enhanced knowledge graph representation model in hyperbolic space. In J. Jiang & W. Chen (Eds.), *Advanced data mining and applications.* Springer. https://doi.org/10.1007/978-3-030-95408-6_11

Xiao, Y., Li, C., & Thürer, M. (2023). A patent recommendation method based on KG repres-entation learning. *Engineering Applications of Artificial Intelligence, 126*, 106722. https://doi.org/10.1016/j.engappai.2023.106722

Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., Xu, B., & Xiao, Y. (2017). CN-DBpedia: A never-ending Chinese knowledge extraction system. In S. Benferhat, K. Tabia, & M. Ali (Eds.), *Advances in artificial intelligence: From theory to practice.* Springer.

Yamauchi, I., & Nagaoka, S. (2015). Does the outsourcing of prior art search increase the efficiency of patent examination? *Evidence from Japan. Research Policy, 44*(8), 1601–1614. https://doi.org/10.1016/j.respol.2015.05.003

Yin, M. J., Wang, B., & Ling, C. (2024). A fast local citation recommendation algorithm scal-able to multi-topics. *Expert Systems with Applications, 238*, 122031. https://doi.org/10.1016/j.eswa.2023.122031

Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics, 10*(4), 1108–1130. https://doi.org/10.1016/j.joi.2016.09.006

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Du, X. (2019). UER: An open-source toolkit for pre-training models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): System demonstrations* (pp. 241–246). https://doi.org/10.18653/v1/D19-3041

Zhao, Y., & Wen, T. (2017). Motivation analysis of patent citation. *Information Studies Theory & Application, 40*(7), 28–32.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.