Information Processing and Management 61 (2024) 103729

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Event is more valuable than you think: Improving the Similar Legal Case Retrieval via event knowledge

Yuxin Zhang [a,b], Songlin Zhai [a,b], Yuan Meng [a,b], Sheng Bi [c], Yongrui Chen [a,b], Guilin Qi [a,b,*]

[a] School of Computer Science and Engineering, Southeast University, No. 2 SEU Road, Nanjing, Jiangsu Province, 211189, China
[b] Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, Nanjing, Jiangsu Province, China
[c] School of Law, Southeast University, No. 2 SEU Road, Nanjing, Jiangsu Province, 211189, China

ARTICLE INFO

Dataset link: https://github.com/YuxinZhangGit/EMV

Keywords:
Event graph
Knowledge graph
Representation learning
Similar legal case retrieval
Natural language processing

ABSTRACT

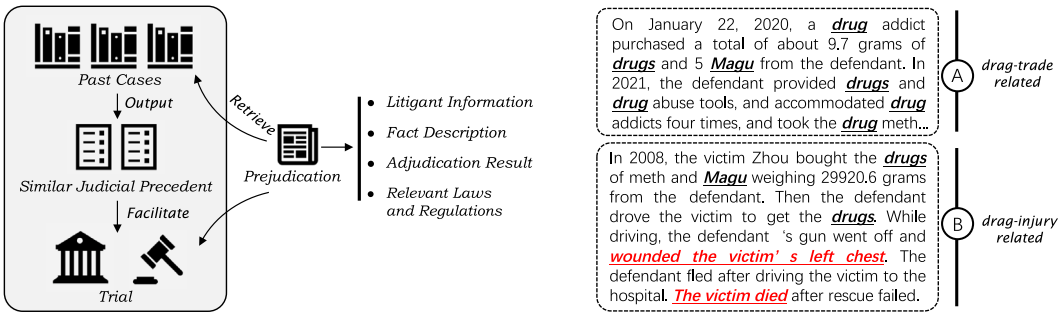
The task of Similar Legal Case Retrieval (SLCR) plays a crucial role in advancing legal assistance systems and enhancing the fairness of judicial outcomes. However, existing methodologies predominantly focus on capturing superficial entity information, largely neglecting the intricate layers of event-related knowledge. This limitation leads to sub-optimal retrieval performance, as cases involving the same entities may diverge substantially in their underlying events, thus yielding inconsistent search results. To bridge this research gap, this paper presents a novel SLCR model, dubbed EMV that enhances the retrieval framework by seamlessly integrating event-based representations. Specifically, EMV employs a heterogeneous knowledge graph designed to facilitate nuanced interactions between entity and event information through a meta-path aggregation mechanism. In addition, the EMV model constructs a comprehensive representation of case features by deeply fusing entity and event information in legal cases and combining the multi-layered characteristics of heterogeneous graphs. This comprehensive information fusion approach not only enhances the model's ability to understand complex information in legal documents but also significantly improves the accuracy and efficiency of legal case retrieval. Experimental results across three public benchmark datasets unequivocally highlight the substantial enhancement brought about by incorporating event knowledge. For example, in tests conducted on the LeCaRD2K dataset, EMV achieved a significant 4.3% increase in accuracy over the current best baseline model, along with a 19.05% decrease in the Mean Rank assessment metric. The superior results indicate that EMV could effectively capture the interrelations between multi-level information within legal cases, thereby establishing a new state-of-the-art standard in the SLCR domain. The datasets and codes are released at https://github.com/YuxinZhangGit/EMV.

1. Introduction

Similar legal case retrieval (SLCR) is an indispensable component of the legal assistance system, serving a crucial role in promoting the fairness of the judicial system. However, manually retrieving is labor-intensive and time-consuming, spurring considerable attention in automatic searching methods (Bench-Capon, Araszkiewicz, Ashley, Atkinson, Bex, Borges, Bourcier,

* Corresponding author. Address: No. 2 SEU Road, Nanjing, Jiangsu Province, 211189, China.
E-mail addresses: zzyx_cs@seu.edu.cn (Y. Zhang), songlin_zhai@seu.edu.cn (S. Zhai), yuan_meng@seu.edu.cn (Y. Meng), bisheng@seu.edu.cn (S. Bi), yrchen@seu.edu.cn (Y. Chen), gqi@seu.edu.cn (G. Qi).

https://doi.org/10.1016/j.ipm.2024.103729
Received 14 November 2023; Received in revised form 31 January 2024; Accepted 23 March 2024
Available online 6 May 2024



(a) Introduction of the SLCR task. (b) Explanation of the importance of event graph.

Fig. 1. The introduction of the SLCR task and the importance of event graph for SLCR.

Bourgine, Conrad, Francesconi, Gordon, Governatori, Leidner, Lewis, Loui, McCarty, Prakken, Schilder, Schweighofer, Thompson, Tyrrell, Verheij, Walton, & Wyner, 2012; Cai, He, Guan, & Li, 2018; Cai & Zheng, 2020; van Opijnen & Santos, 2017). As depicted in Fig. 1(a), the primary objective of this task is to identify relevant judicial precedents that could provide guidance and enhance the credibility and fairness of current trials (Chen, Wu, Chen, Lu, & Ding, 2022). This paper will also follow this research line, aiming to improve the SLCR task.

Different from the matching of semantic similarity in traditional textual matching research, the SLCR task needs to explore the deeper legal-related logical relation of legal cases beyond the surface semantics (Hu et al., 2022; Shao et al., 2020), which remains a significant challenging endeavor due to the inherent intricacy of legal documents. Specifically, as shown in Fig. 1(a), a typical legal case comprises four fundamental parts (Bench-Capon et al., 2012; Hu et al., 2022): litigant information, fact description, adjudication result, and relevant laws and regulations. Among these, the "fact description" acts as a representative feature in legal documents as it provides a comprehensive account of the case, *e.g.,* the circumstances and elements of the crime. This description is generally employed as the yardstick to measure the similarity between different legal cases. However, due to the intricate legal case structure, effectively identifying nuanced differences within these descriptions still poses a significant challenge (Lyu et al., 2022; Zhao, Gao, & Guo, 2023). Previous research has primarily contributed to resolving this issue from two directions: (1) the utilization of explicit textual statistical features (Trivedi, Trivedi, Varshney, Joshipura, Mehta, & Dhanani, 2021) via topic models (Lu & Conrad, 2012) or TF-IDF methods (Kumar, Reddy, Reddy, & Suri, 2013); (2) the adoption of contextualized semantic features based on pre-trained language models (PLMs) (Shao et al., 2020; Wu et al., 2020). While some researchers also attempt to incorporate legal features to enhance the model comprehension of legal documents, these methods remain confined to statistical or semantic abstractions at the entity level and fail to consider the more complex layers of event-based knowledge (Bhattacharya, Ghosh, Pal, & Ghosh, 2022). Consequently, the state-of-the-art in SLCR is yet to achieve optimal performance.

To enhance the effectiveness of the legal case retrieval task, this paper primarily focuses on measuring the factual similarity between legal documents based on the event graph. In this context, each node (*i.e.,* event) in the graph generally refers to an action and its related attributes (Lou, Liao, Deng, Zhang, & Chen, 2021), *e.g.,* the subject, object, or mediator of the action, with the "trigger" describing the action and "arguments" for attributes (Guan et al., 2023). This event knowledge plays a crucial role in providing a comprehensive depiction of legal facts, significantly influencing the model performance (Song, Li, Cai, Yang, Yang, & Liu, 2022). To illustrate this concept, Fig. 1(b) depicts two real-world legal cases that share a similar entity (*i.e.,* drug) but involve different events (*i.e.,* drag-trade and drag-injury). The distinct event leads to different retrieval results for the judicial precedent. This observation highlights the potential bias introduced by high-frequency entities and emphasizes the advantages of incorporating event information to mitigate the risk of misjudgment (Sun, Huang, Xu, Chen, Ren, & Hu, 2023).

In this paper, we propose a novel model to address this issue, dubbed `EMV` (*i.e.,* the acronym of *Event is More Valuable*), where the event knowledge is also integrated into a heterogeneous knowledge graph to enable more accurate retrieval of similar legal cases. This heterogeneous graph incorporates not only the crucial event graph but also the entity information, allowing the legal case could be represented from multiple perspectives. Specifically, the event graph focuses on the event sequences to summarize legal facts and assists the model in extracting relevant features, while the entity graph utilizes the external legal knowledge base `LegalKG` (Bi, Huang, Cheng, Wang, & Qi, 2019) to help the model understand the semantics of terms in the legal field. Combining the event and entity graphs, the heterogeneous knowledge graph enhances the model's ability to identify subtle differences between legal cases by considering their interactions. To validate the `EMV`'s effectiveness, we conduct extensive experiments on various legal case datasets. The experimental results demonstrate that the proposed `EMV` model outperforms the state-of-the-art model by a large margin of 4.3% in this SLCR task.

To summarize, the contributions of this paper could be listed as follows:

- This paper emphasizes the significance of event knowledge in improving the effectiveness of the SLCR task and pioneers the integration of event information as a novel methodology in this context.

- A novel model (dubbed EMV) is presented in this paper that aims to incorporate the crucial event graph information. Additionally, a heterogeneous graph is also introduced to consider both event and entity knowledge and their interaction information through the meta-path aggregation algorithm.
- Extensive experiments are conducted on three public benchmark datasets to validate the effectiveness of the proposed EMV model. The experimental results demonstrate that EMV outperforms existing models, establishing it as the state-of-the-art approach for the SLCR task across various datasets.

2. Related work

Similar legal case retrieval is a challenging task, and how to judge the similarity between legal cases is a problem worth exploring. Initial methods of determining similar legal cases rely on statistical results (Trivedi et al., 2021). Winkels, Boer, Vredebregt, and van Someren (2014) utilized a topic model to assign different topics to legal documents and performed similar legal case matching through topic similarity, while Kumar et al. (2013) employed TF-IDF to collect term frequency-inverse document frequency information of words in different documents to generalize document-level features. Bhattacharya, Ghosh, Pal, and Ghosh (2020) measured legal document similarity by mining features of citation information between legal documents. However, statistical-based methods concentrate on explicit textual features but fail to capture underlying semantic information.

Some research works utilized pre-trained models to embed documents into a low-dimensional continuous vector space and measured the relevance of legal documents through semantic similarity. Hong, Zhou, Zhang, Li, and Mo (2020) proposed the Legal Feature Enhanced Semantic Matching Network (LFESM) to capture subtle distinctions between legal document pairs. Shao et al. (2020) proposed a legal case retrieval model BERT-PLI, which employed a pre-trained model to encode paragraph-level semantics and aggregated paragraph-level interaction information to deduce document relevance. In order to augment understanding of Chinese legal long-text documents, Xiao, Hu, Liu, Tu, and Sun (2021) proposed a pre-trained language model Lawformer, which can handle long texts.

Actually, some problems exist in similar Chinese legal case retrieval methods based on pre-trained language models. Legal documents contain a large amount of domain-specific knowledge, which is difficult for those methods based on pre-trained models to understand. Li, Lu, Le, and He (2022) proposed the Interactive Attention Capsule Network (IACN), performing interpretable predictions by capturing fine-grained element-level similarity. Bi et al. (2019) constructed a hybrid heterogeneous graph containing term network and external knowledge through legal documents and legal encyclopedia data and utilized this graph to capture the intrinsic relationship between legal documents and legal domain knowledge.

However, the above approaches failed to consider event-level information in legal documents. In addition, some researchers point out that high-frequency entity information may cause model bias. In contrast with discrete entity-level knowledge, event-level information summarizes the complete course of criminal facts in legal judgments. Therefore, we attempt to integrate the event knowledge to achieve fine-grained similar case retrieval.

3. Methodology

3.1. Notations

For the legal document set $C = \{c_k \mid 1 \le k \le N\}$, it denotes the collection of legal cases to be retrieved. To obtain a better case representation and facilitate the retrieving, each document c_k is used to construct an event sub-graph g_k^1, an entity sub-graph g_k^2, and a heterogeneous graph g_k. As a result, this legal document set could form an event graph set $\mathcal{G}^1$, an entity graph set $\mathcal{G}^2$ and a heterogeneous graph set $\mathcal{G}$, which will be introduced next.

Heterogeneous Graph: A heterogeneous graph g_k is also constructed from each document c_k to model the event knowledge comprehensively (Liu, Wu, Liu, & Qian, 2023). For one graph g_k in the heterogeneous graph set $\mathcal{G} = \{g_k \mid 1 \le k \le N\}$, it is also comprised of a vertex set $\mathcal{V} = \{v_i \mid 1 \le i \le |\mathcal{V}|\}$ and edge set $\mathcal{E} = \{e_j \mid 1 \le j \le |\mathcal{E}|\}$. For the vertex set, it is the combination of the event and entity vertex sets, *i.e.,* $\mathcal{V} = \mathcal{V}^1 \cup \mathcal{V}^2$. For each edge $e_i \in \mathcal{E}$ in g_k, it is one of these three types, including "*event-event*", "*event-entity*" or "*entity-entity*", which means that the pair of nodes connected by this edge (*e.g.,* $(v_j, v_k) = e_i$) might indicate a different type of nodes. Furthermore, the path formed by the edges in g_k could also be treated as the instance of three pre-defined meta-path $\{\pi, \mu, \tau\}$, where π denotes *event-entity* with μ for *event-entity-entity* and τ for *event-event-entity*. These meta-paths denote the different types of influencing processes of a given event on other nodes. Specifically, $\pi(v_i)$ (or $\mu(v_i)$, $\tau(v_i)$) indicates the instance set of the meta-path π (or μ, τ), where each instance refers to a path starting from the node v_i.

Event Graph: For an event graph in the graph set $\mathcal{G}^1 = \{g_k^1 \mid 1 \le k \le N\}$, it could be denoted as an event-vertex set $\mathcal{V}^1 = \{v_i^1 \mid 1 \le i \le |\mathcal{V}^1|\}$ and an edge set $\mathcal{E}^1 = \{e_j^1 \mid 1 \le j \le |\mathcal{E}^1|\}$, *i.e.,* $g_k^1 = \{\mathcal{V}^1, \mathcal{E}^1\}$. Specifically, $\mathcal{N}_i^1$ indicates the neighbor node set of event node v_i^1 in g_k^1. Additionally, it should be noted that there are two types of edges between events, including the time-sequential and the causal edges, thereby the edge set $\mathcal{E}^1$ could also be denoted as $\mathcal{E}^1 = \{\mathcal{E}^t, \mathcal{E}^c\}$ with $\mathcal{E}^t$ and $\mathcal{E}^c$ being the time-sequential and causal sets respectively.

Entity Graph: Analogously, an entity graph g_k^2 in the graph set $\mathcal{G}^2 = \{g_k^2 \mid 1 \le k \le N\}$, it could be also represented as the entity set $\mathcal{V}^2 = \{v_i^2 \mid 1 \le i \le |\mathcal{V}^2|\}$ and its edge set $\mathcal{E}^2 = \{e_j^2 \mid 1 \le j \le |\mathcal{E}^2|\}$, *i.e.,* $g_k^2 = \{\mathcal{V}^2, \mathcal{E}^2\}$. Additionally, $\mathcal{N}_i^2$ denotes the neighbor node set of entity v_i^2 in g_k^2.

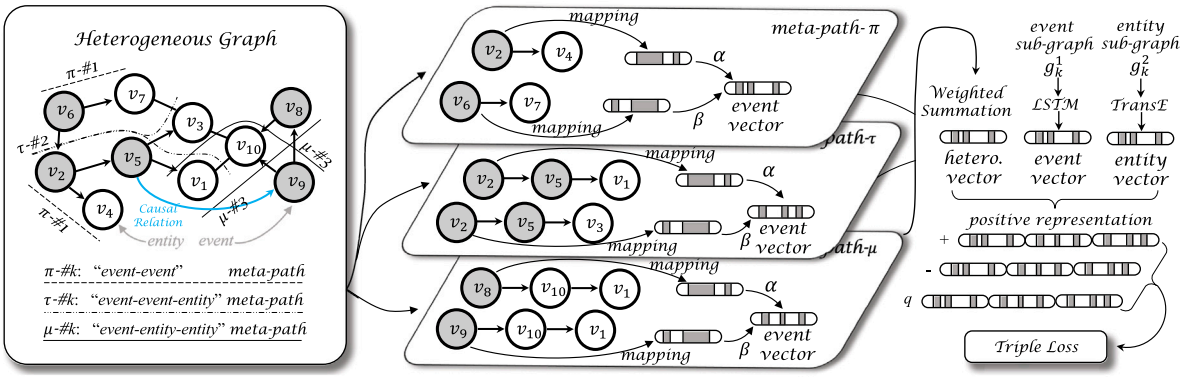
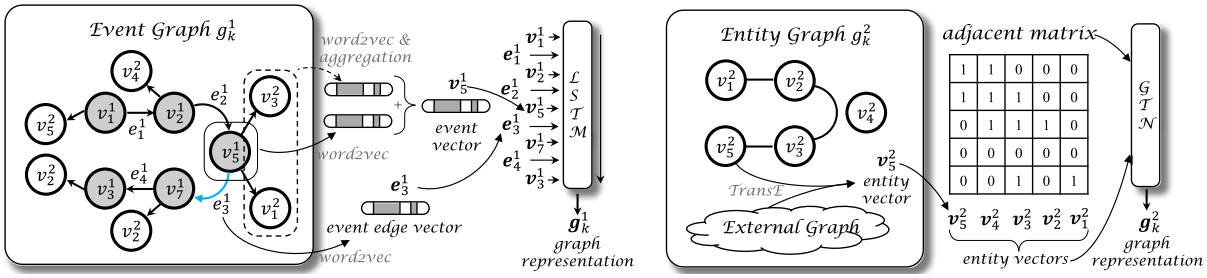


Fig. 2. The architecture of the EMV framework, where the gray and white vertices in the heterogeneous graph signify the events and entities, respectively. Here, event knowledge is comprehensively encapsulated using both event and heterogeneous graphs, bolstering the representation of legal documents. Concurrently, entity knowledge within the legal document is seamlessly integrated, enriching the overall representation.



(a) Representation of event graph.

(b) Explanation of the importance of entity graph.

Fig. 3. The left figure demonstrates how EMV synergizes event coding with sequence models to derive the event graph's representation. Conversely, the right one clarifies that EMV harnesses both TransE and GTN to learn the representation of the entity graph.

3.2. Event knowledge in event graph

The representation of an event graph needs to incorporate not only the latent semantic information but also the sequential information of events. To achieve this target, EMV employs a sequence model to encode the event-level semantic information, where the multiple events could form the event chain connected by the relations between events. Formally, the semantic feature of the event v_i^1 could be formulated as:

$$\boldsymbol{v}_i^1 = \phi(v_i^1) \oplus \sum_{j=1}^{|\mathcal{N}_i^1|} \phi(v_j^1) \tag{1}$$

where $\boldsymbol{v}_i^1$ is the representation of event v_i^1. ϕ denotes the word2vec operation based on the event trigger word. $\mathcal{N}_i^1$ represents the set of adjacent nodes of v_i^1, and $\oplus$ stands for concatenation operator. Analogously, the edge (*e.g.,* e_i^1) in the event graph could also be represented as:

$$\boldsymbol{e}_i^1 = \phi(e_i^1) \tag{2}$$

where $\phi(e_i^1)$ generates the edge embedding vector based on the edge trigger words.

Based on Eqs. (1) and (2), the event chain contained in the event graph could be represented as a sequence of feature vectors, *e.g.,* $\{\boldsymbol{v}_1^1, \boldsymbol{e}_1^1, \boldsymbol{v}_2^1, \ldots, \boldsymbol{e}_{|\mathcal{E}^1|}^1, \boldsymbol{v}_{|\mathcal{V}^1|}^1\}$. In order to effectively mine the semantics of this sequence in the event knowledge network, we apply an LSTM to encode this event chain. The corresponding initial vector sequence is taken as the input, and the final output hidden variable represents the embedding of the event graph, *e.g.,* $\boldsymbol{g}^1$:

$$\boldsymbol{g}^1 = \text{LSTM}(\{\boldsymbol{v}_1^1, \boldsymbol{e}_1^1, \boldsymbol{v}_2^1, \ldots, \boldsymbol{e}_{|\mathcal{E}^1|}^1, \boldsymbol{v}_{|\mathcal{V}^1|}^1\}) \tag{3}$$

This process has been illustrated in Fig. 3(a). Notably, the encoding method of the event chain can be flexibly extended, and various sequence feature encoding methods could also be adapted to this task, such as RNN or Transformer Encoder.

3.3. Entity knowledge in entity graph

Entity representation needs to fully consider the relations between entities. There are only simple correlations between entities in the entity knowledge network we constructed. Given an entity graph $g_k^2 = \{\mathcal{V}^2, \mathcal{E}^2\}$, and the entity knowledge encoder represents the projection function that learns the graph g_k^2 to $\mathbf{g}_k^2$. To begin, a placeholder node is added to the entity graph as a global embedding node. EMV leveraging the TransE knowledge graph embedding technique to initialize the feature vectors $\mathcal{X}_k$ for entity nodes. Subsequently, the adjacency matrix $\mathcal{A}_k$, in tandem with the entity graphs' feature vectors, is fed into the Graph Transformer Network (Yun, Jeong, Kim, Kang, & Kim, 2019) (GTN). This encoder adeptly captures the semantic nuances and intricate interactions between entities by adopting a self-attention mechanism.

$$\mathbf{g}_k^2 = \mathrm{GTN}\{\mathcal{A}_k, \mathcal{X}_k\} \tag{4}$$

Ultimately, the representation of the entity graph's knowledge is the result of encoding the global embedding nodes. The process has been also depicted in Fig. 3(b).

3.4. Event knowledge in heterogeneous graph

In the heterogeneous graph, there are three different types of meta-paths. We first learn the event representation under three different meta-paths to incorporate the event knowledge in the heterogeneous graph. Different meta-path instances start from this node for an event node in this graph. Specifically, taking the τ-meta-path starting from $v_2 \in \mathcal{V} \cup \mathcal{V}^1$ (*i.e.,* $\tau(v_2)$) as an example, it can be instanced as different paths, *e.g.,* $v_2 \xrightarrow{e_1} v_5 \xrightarrow{e_2} v_1$ or $v_2 \xrightarrow{e_1} v_5 \xrightarrow{e_2} v_3$, which have been shown in Fig. 2. According to Eqs. (1) and (2), a given instance could be represented as a sequence of the nodes and edges representations (*e.g.,* $\{\boldsymbol{v}_2, \boldsymbol{e}_1, \boldsymbol{v}_5, \boldsymbol{e}_2, \boldsymbol{v}_1\}$). This instance describes the change of the event node v_2, which motivates us to represent this event node based on the instance dynamics. To be specific, we add the time stamp into the path for describing the changes from the event, *i.e.,* $\{\boldsymbol{v}_{t=1} \xrightarrow{\boldsymbol{e}_{t=1}} \boldsymbol{v}_{t=2} \xrightarrow{\boldsymbol{e}_{t=2}} \boldsymbol{v}_{t=3}\}$, thereby formulating the event node v_2 as:

$$\boldsymbol{h}_t = I(\boldsymbol{v}_t) \cdot \boldsymbol{v}_t + \boldsymbol{h}_{t-1} \odot I(\boldsymbol{e}_{t-1}) \cdot \boldsymbol{e}_{t-1}, \quad 1 \le t \le 2 \tag{5}$$

where $\boldsymbol{h}_t$ denotes the hidden state at time t with being a zero vector at $t = 0$. $\odot$ refers to the Hadamard product. $I(\cdot)$ is a piece-wise function that will return the corresponding transformation matrices based on its inputs to map the node or edge representations into a unified space:

$$I(X) = \begin{cases} \boldsymbol{W}^1, & \text{if} X \in \mathcal{V}^1 \quad \#\ X \text{ being the event node;} \\ \boldsymbol{W}^2, & \text{if} X \in \mathcal{V}^2 \quad \#\ X \text{ being the entity node;} \\ \boldsymbol{W}^t, & \text{if} X \in \mathcal{E}^t \quad \#\ X \text{ being the time-sequential edge;} \\ \boldsymbol{W}^c, & \text{if} X \in \mathcal{E}^c \quad \#\ X \text{ being the causal edge;} \\ \boldsymbol{W}^e, & \text{if} X \in \mathcal{E}^2 \quad \#\ X \text{ being the entity edge} \end{cases} \tag{6}$$

where $\boldsymbol{W}^1$ and $\boldsymbol{W}^2$ are used to map the different event and entity nodes in the heterogeneous graph, respectively. Analogously, $\boldsymbol{W}^t$, $\boldsymbol{W}^c$, and $\boldsymbol{W}^e$ denote the mapping matrices of the time-sequential, causal, and entity edge. As such, the representation of one event node v_i under its $j-$th τ-meta-path instance could be formulated as the mean of the hidden states estimated by Eq. (5):

$$\boldsymbol{v}_{i,j}^{\tau} = \sum_t \boldsymbol{h}_t / (T_j + 1) \tag{7}$$

where T_j denotes the length of dynamic list of the jth instance in the τ-meta-path set $\tau(v_i)$.

By applying Eq. (7), we can obtain the different representations of an event node v_i under different instances of the τ-meta-path. With these features in hand, the node v_i could be represented as the weighted summation of them:

$$\boldsymbol{v}_i^{\tau} = \mathrm{ReLu}\Big(\sum_{j=1}^{|\tau(v_i)|} \alpha_j \boldsymbol{v}_{i,j}^{\tau}\Big) \tag{8}$$

where $\boldsymbol{v}_i^{\tau}$ denotes the representation of node v_i under the τ-meta-path with $\boldsymbol{v}_{i,j}^{\tau}$ being the representation of v_i under jth instance of τ-meta-path that starts from it in the heterogeneous graph. $\tau(v_j)$ is the set of τ-meta-path that starts from v_j in the heterogeneous graph, with $|\tau(v_j)|$ being its size. α_j is the weight, calculated as:

$$\alpha_j = e^{z_j} \Big/ \sum_{j=1}^{|\tau(v_i)|} e^{z_j}, \quad z_j = \mathrm{LeakyReLu}((\boldsymbol{a}^{\tau})^{\top} \cdot ((I(v_i) \cdot \boldsymbol{v}_i) \oplus \boldsymbol{v}_{i,j}^{\tau})) \tag{9}$$

where e denotes the exponents of natural numbers. $\boldsymbol{a}^{\tau}$ is the learnable parameter vector for the τ-meta-path. For other types of meta-paths, the corresponding vectors will also be provided, *i.e.,* $\boldsymbol{a}^{\pi}$ and $\boldsymbol{a}^{\mu}$.

Analogously, the node v_i could also be represented under the π-meta-path and μ-meta-path, with their representations being $\boldsymbol{v}_i^{\pi}$ and $\boldsymbol{v}_i^{\mu}$. As such, the event node in the heterogeneous graph could be formulated by the weighted summation of them:

$$\boldsymbol{v}_i = \beta_1 \boldsymbol{v}_i^{\tau} + \beta_2 \boldsymbol{v}_i^{\pi} + \beta_3 \boldsymbol{v}_i^{\mu} \tag{10}$$

Table 1
Statistics of three benchmark datasets in this paper, where #Triplet indicates the number of the triplet and ¯ refers to the average value.

Dataset	#Triplet (Train)	#Triplet (Test)	$\overline{\#\text{Entity}}$	$\overline{\text{Document Length}}$
CAIL2019SCM	7171	1793	22.15	442.38
LeCaRD2K	1600	400	17.48	355.92
LDC2K	1600	400	20.60	382.26

where $\beta_{\cdot}$ denotes the weight of the corresponding meta-path representation. Take the weight of τ-meta-path β_1 as an example, and it could be estimated as:

$$\beta_1 = e^{z_1} / \sum_{j=1}^{3} e^{z_j}, \quad \text{for } j = 1,\ z_1 = (\boldsymbol{b}^{\tau})^{\top} \cdot \left(\frac{1}{|\mathcal{V}^1|} \sum_{i=1}^{|\mathcal{V}^1|} \tanh(\boldsymbol{W}^{\tau} \cdot \boldsymbol{v}_i^{\tau} + \boldsymbol{\epsilon}^{\tau})\right) \tag{11}$$

where e denotes the exponents of natural numbers. $\boldsymbol{b}^{\tau}$ is the learnable parameter vector for the τ-meta-path. $\boldsymbol{W}^{\tau}$ and $\boldsymbol{\epsilon}^{\tau}$ denote the transformation matrix and error vector for the τ-meta-path, respectively. For other types of meta-paths, the corresponding transformation matrices and error vectors are also provided, *i.e.,* $\boldsymbol{W}^{\pi}$, $\boldsymbol{W}^{\mu}$ and $\boldsymbol{\epsilon}^{\pi}$ and $\boldsymbol{\epsilon}^{\mu}$.

Equipped with the node representation shown in Eq. (10), the heterogeneous knowledge graph (*e.g.,* g_k) could be represented by encoding the sequence information of the event chain in this graph based on LSTM:

$$\boldsymbol{g}_k = \texttt{LSTM}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_{|\mathcal{V}^1|}\}), \quad v_i \in \mathcal{V} \cap \mathcal{V}^1 \tag{12}$$

3.5. Event knowledge fusion for legal document

As shown in Eqs. (3) and (12), the event knowledge could be fully represented from the event and heterogeneous graphs, which support the representation of the legal documents. Additionally, the textual semantics and entity knowledge in the legal document are also fused to further enrich its representation, formulated as:

$$\boldsymbol{c}_k = \tanh(\boldsymbol{M} \cdot \boldsymbol{g}_k + \boldsymbol{\epsilon}) \oplus \tanh(\boldsymbol{M}^1 \cdot \boldsymbol{g}_k^1 + \boldsymbol{\epsilon}^1) \oplus \tanh(\boldsymbol{M}^2 \cdot \boldsymbol{g}_k^2 + \boldsymbol{\epsilon}^2) \oplus \tanh(\boldsymbol{M}^3 \cdot \boldsymbol{c}_k' + \boldsymbol{\epsilon}^3). \tag{13}$$

This representation fusion module fuses text representation, entity knowledge network representation, event knowledge network representation, and heterogeneous knowledge network representation. $\boldsymbol{M}^{\cdot}$ is the transformation matrix to ensure that the representations from different sources are in the unified feature space. $\boldsymbol{\epsilon}^{\cdot}$ denotes the learnable noise vector. $\boldsymbol{c}_k'$ indicates the textual representation of the legal document c_k, learned by the pre-trained language model.

3.6. Training

EMV calculates the loss function based on the pairwise mode in the model training phase and optimizes the ranking problem by similarity. Each training dataset are organized as triplets (c_k, c_k^+, c_k^-) with the representations being $(\boldsymbol{c}_k, \boldsymbol{c}_k^+, \boldsymbol{c}_k^-)$, where c_k is the query legal text with c_k^+ and c_k^- being its positive and negative documents. The EMV's training process is formulated as follows:

$$\mathcal{L}_{\Theta}(C) = \sum_{c_k \in C} \max(0, [\gamma - \text{sim}(\boldsymbol{c}_k, \boldsymbol{c}_k^+) + \text{sim}(\boldsymbol{c}_k, \boldsymbol{c}_k^-)]) \tag{14}$$

where γ is margin, and sim$(\cdot)$ represents the similarity calculation function between feature vectors, implemented by the Cosine similarity in this paper. Θ is the model parameters that need to be learned.

4. Experiment

4.1. Experimental setup

4.1.1. Datasets

We evaluate our model and all compared baselines on three Chinese similar legal case datasets, including CAIL2019-SCM, LeCaRD2K, and LDC2K:

- **CAIL2019SCM** originates from the similar case matching task hosted within the *Challenge of AI in Law*, as organized by Xiao et al. (2019). It comprises a total of 8964 similar case triplets, each consisting of one inquiry legal text and two candidate legal texts.
- **LeCaRD2K** is another Chinese legal case retrieval dataset introduced by Ma et al. (2021). This dataset comprises 107 legal query documents, each accompanied by a set of 100 candidate legal documents. Specifically, we conducted a meticulous manual curation process to construct 2000 high-quality similar case triplets within LeCaRD2K. This construction involved a synthesis of expert insights and adherence to rigorous evaluation criteria for legal similarity.

- **LDC2K**: We collect 271,295 legal documents in seven common charges from China Judgments Online[1] (Bi, Ali, Wang, Wu, & Qi, 2022), and conduct preliminary screening and structured processing of each legal document. In addition, we manually constructed 2000 high-quality similar case triplets based on key elements of party information and judgment results, combined with document TF-IDF similarity and legal expert opinions.

The detailed statistical information for the three datasets is presented in Table 1. The LeCaRD2K and LDC2K datasets consist of manually constructed similar case triplets, thus comprising only 2000 triplets each. The CAIL2019SCM dataset contains 8964 similar case triplets. We divided the datasets in a $2:8$ ratio to serve as test and training sets, respectively. Among these, the most complex dataset, CAIL2019SCM, features an average of 22.15 entities and 442.38 tokens per legal document.

4.1.2. Knowledge graph construction

Consider that legal knowledge graph construction is cumbersome, and this task is not the focus of this research. Therefore, this paper does not delve into the intricacies of the related work. We follow the legal knowledge graph `LegalKG` proposed by Bi et al. (2019). They extract entities from related texts such as encyclopedias and legal documents and construct related relations according to the co-occurrence frequency of entities. Given a piece of text c_k, we extract a set of legal-related entities $\mathcal{E}^2$ from the text and then perform entity alignment with `LegalKG`. By connecting the entities in c_k according to the relations between entities in `LegalKG`, we can construct the entity graph g_k^2 for the legal text.

The construction of an event graph is a two-phase process involving event extraction and relation extraction. Initially, we utilize `PAJHEE` (Shen, Qi, Li, Bi, & Wang, 2020) to extract events in the "event description" of legal documents, *i.e.,* event trigger words, event types, and arguments. Subsequently, we employ an event-causal extraction method based on derived prompt learning to extract causal and temporal relations between events (Shen, Zhou, Wu, & Qi, 2022). The event graph is represented as $g_k^1 = (\mathcal{V}^1, \mathcal{E}^1)$, where $\mathcal{E}^1 = \mathcal{E}^c, \mathcal{E}^t$ represents the set of causal and temporal edges.

Inspired by the legal hybrid knowledge network (Bi et al., 2022), we further construct a heterogeneous knowledge graph by aligning event arguments and entities. Our initial step entails matching the vertices representing event arguments in g_k^1 with those representing legal entities in g_k^1. After this matching phase, the aligned vertices are amalgamated into a singular vertex. Concurrently, vertices that do not align, including both argument and legal entity vertices, are discarded. Ultimately, we obtain a heterogeneous graph g_k that integrates event and entity information.

4.1.3. Evaluation metrics

To ascertain the model's prowess in retrieving similar legal cases, we employ two pivotal evaluation metrics:

- **ACC**: Given a set of similar case triplets (c_i, c_i^+, c_i^-), where c_i denotes the query text and c_i^+, c_i^- are the candidate texts, the objective of similar legal case matching is to discern the candidate text bearing a higher resemblance to the query:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\text{sim}(c_i, c_i^+) > \text{sim}(c_i, c_i^-)\right) \tag{15}$$

  where $\mathbb{I}$ indicates the indicator function, N is the total number of similar case triplets. $\text{sim}(\cdot,\cdot)$ refers to the `cosine` similarity based on the legal document representations extracted according to Eq. (13).
- **MR** offers an evaluation paradigm that mirrors real-world applications more closely, thus providing a more authentic reflection of the model's operational performance. For a given similar case pair (c_i, c_i^+), the evaluation metric Mean Rank (MR) captures the relative position of the ground truth c_i^+ amidst all candidate texts, estimated by:

$$\text{MR} = \frac{1}{M}\sum_{i=1}^{M}\text{Rank}(c_i^+) \tag{16}$$

  where $\text{Rank}(c_i^+)$ returns the position of the ground truth c_i^+ for the given similar case pair (c_i, c_i^+) among all candidate texts, M is the total number of similar case pairs.

4.1.4. Baselines

This section introduces some of the most popular and state-of-the-art methods in text-matching tasks as baselines. To comprehensively assess the effectiveness of our proposed approach, we conduct a comparative analysis with several state-of-the-art methods in the domain of text matching:

- **BERT** (Devlin, Chang, Lee, & Toutanova, 2019) is a pre-trained language model with bidirectionally encoded representations. In particular, we select the BERT trained on the legal corpus released by Zhang, Zhang, Liu and Sun (2019).
- **RoBERTa** (Liu et al., 2019) improves the pre-training task designed based on BERT and is trained with larger batches on a larger corpus.
- **T5** (Raffel et al., 2020) is a pre-trained model based on the Transformer architecture that employs a unified "text-to-text" format when dealing with various natural language processing tasks.

[1] https://wenshu.court.gov.cn/

Table 2
Performance metrics for similar case retrieval across diverse models on the CAIL2019SCM, LeCaRD2K, and LDC2K datasets. Results are quantified using Acc. (Accuracy) and MR (Mean Rank). Entries in bold represent the best result in each category.

Model	CAIL2019SCM		LeCaRD2K		LDC2K	
	Acc.	MR	Acc.	MR	Acc.	MR
BERT-base	71.29	311.37	73.03	89.66	72.86	99.35
RoBERTa-base	71.79	296.63	73.21	80.14	72.84	83.97
RoBERTa-large	72.26	289.47	74.62	83.81	73.92	80.22
T5-base	73.45	303.94	79.67	78.77	75.19	86.25
T5-large	74.84	299.41	79.10	74.09	75.27	83.27
GPT2	74.31	283.40	79.29	72.17	76.87	89.85
KALM	73.97	176.03	77.93	46.76	75.81	44.50
KnowBERT	71.96	187.20	76.44	50.54	75.88	48.57
Lawformer	72.36	208.96	74.84	68.62	72.28	66.18
LFESM	72.35	224.80	74.20	85.27	72.61	82.93
AlphaCourt	72.44	355.42	71.46	107.16	72.20	108.80
HetGRL	73.18	–	–	–	–	–
TextCNN	67.40	486.18	69.85	142.57	67.24	157.83
TextRNN	68.77	469.65	70.80	136.49	68.43	134.96
EMV	**75.70** (±0.6)	**145.55** (±7.3)	**81.30** (±0.8)	**37.85** (±2.8)	**77.90** (±0.3)	**35.27** (±3.6)

- **GPT2** (Alec, Jeffrey, Rewon, David, Dario, & Ilya, 2019) is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages.
- **KALM** (Feng, Tan, Zhang, Lei, & Tsvetkov, 2023) is a Knowledge-Aware Language Model that jointly leverages knowledge in local, document-level, and global contexts for long document understanding.
- **KnowBERT** (Peters, Neumann, IV, Schwartz, Joshi, Singh, & Smith, 2019) uses an integrated entity linker to retrieve relevant entity embeddings, and then update contextual word representations via a form of word-to-entity attention.
- **Lawformer** (Xiao et al., 2021) is a long-text pre-trained model based on Longformer to introduce legal domain knowledge.
- **AlphaCourt** is the champion solution of the 2019 Challenge of AI in Law SCM task. However, the main body of the model integrates five sub-modules based on pre-trained language models, making the training process very time-consuming.
- **LFESM** (Hong et al., 2020) is a legal feature-enhanced semantic matching network that utilizes a Siamese network framework to capture subtle feature differences between query text and candidate text.
- **HetGRL** (Bi et al., 2022) constructs a legal document-legal entity heterogeneous graph and obtains the semantic representation vector of a legal document that incorporates external legal knowledge through graph encoding.
- **TextCNN** (Kim, 2014) utilizes a convolutional neural network to capture local semantic features of text.
- **TextRNN** (Liu, Qiu, & Huang, 2016) utilizes a recurrent neural network to model text sequence features and mine the underlying semantic information in the text.

4.1.5. Parameter setting

Given the distinct parameter distributions and structural variations among the sub-modules within EMV, we adopt separate hyper-parameters and optimization strategies for each sub-module. For the pre-trained language model of the text representation module, the text embedding vector remains 768-dimensional, and using AdamW optimizer with a learning rate of $2e$-6. In the event knowledge network, entity knowledge network, and heterogeneous network, their representations are set to 256-dimensional, with a learning rate of $5e$-3. Additionally, during the fusion stage, the triplet loss is computed with a margin of 1.0. It is important to note that these hyper-parameters are fine-tuned through grid search. For other hyper-parameters, we maintain the default parameters in the original papers. Specifically, in all our experiments, we report the average results across 5 runs to ensure reliability.

4.2. Performance evaluation

We examine the performance of EMV and other benchmarks on two metrics across three datasets. Table 2 shows the results of the Acc. and MR performance metrics for each model. Generally, we observe that the proposed EMV consistently gains the optimal performance on different datasets and all evaluation metrics, achieving 75.70%, 81.30%, and 77.90% on Acc. across CAIL2019SCM, LeCaRD2K, and LDC2K datasets. Specifically, EMV achieves 145.55, 37.85, and 35.27 mean rank metrics much lower than the state-of-the-art baseline model, indicating superior ranking performance. In addition, EMV has a small error range in all three datasets, *e.g.,* Acc. of ±0.6 on CAIL2019SCM, which suggests that the model's performance is stable and robust.

Regarding to the Acc., it is worth noting that traditional models such as TextCNN and TextRNN reach lower performance on all datasets, suggesting that these methods may not be suitable for handling such tasks compared to deep learning methods. Furthermore, we notice that generative pre-trained language models employed to compare text similarity, *e.g.,* GPT2 and T5, manifesto superior performance over the knowledge-enhanced models (*e.g.,* KALM and KnowBERT).

For the MR metric, we observe that the pre-trained language models (*e.g.,* BERT, RoBERTa) could achieve satisfactory performance, demonstrating that such models simply fit the optimal objective and cannot be adapted to real-world legal document

Table 3
Ablation analysis results of EMV with different heterogeneous knowledge graph neural network representation methods in different datasets. Here, k signifies the number of layers in the graph neural network.

Graph Neural Network		CAIL2019SCM		LeCaRD2K		LDC2K	
		Acc.	MR	Acc.	MR	Acc.	MR
EMV	$k = 2$	74.65	162.96	78.58	46.86	76.18	40.51
	$k = 3$	**75.70** (±0.6)	**145.55** (±7.3)	**81.30** (±0.8)	**37.85** (±2.8)	**77.90** (±0.3)	**35.27** (±3.6)
$-w$ HAN	$k = 2$	75.65	158.82	79.58	50.63	76.37	46.74
	$k = 3$	72.95	182.47	74.10	60.48	73.07	62.33
$-w$ HetGNN	$k = 2$	72.46	180.35	76.89	59.88	74.31	66.29
	$k = 3$	73.98	165.96	77.48	53.00	74.63	52.99
$-w$ GTN	$k = 2$	74.14	176.83	76.81	55.18	74.16	67.52
	$k = 3$	74.88	150.10	78.74	39.34	75.77	43.42

retrieval scenarios. Additionally, we observe that HetGRL performs poorly on MR, indicating that the unsupervised graph-based neural network has poor scalability and could not solve the OOV problem. Models such as KALM and KnowBERT (Knowledge Enhanced Language Models for Fusing Text and Entities) are strong baseline since they are only preceded by EMV, implying that additional domain knowledge can help the model obtain richer semantic information. Besides, EMV that fuses legal entity knowledge and event knowledge outperforms models that only focus on entity knowledge. This finding verifies that both the entity and event knowledge in heterogeneous and different levels of incorporating knowledge information can help EMV learn stronger textual feature representations.

4.3. Analysis

4.3.1. Analysis of the representation of heterogeneous graph

To further verify the effectiveness of the heterogeneous knowledge network representation module based on the meta-path aggregation process, we embed the heterogeneous graph using different representation models and test their retrieval performance:

- **HAN** (Wang et al., 2019) is a heterogeneous graph neural network based on two-layer attention at the node and semantic level, aggregating target node embeddings by fully considering the importance of nodes and meta-paths.
- **HetGNN** (Zhang, Song, Huang, Swami & Chawla, 2019) formalizes the problem of heterogeneous graph representation learning and proposes a heterogeneous graph neural network model, enabling the capture of both structural and content features of heterogeneous networks.
- **GTN** (Yun et al., 2019) learns a soft selection of edge types and composite relations to generate multi-hop connections, where its core component is the graph transformer layer.

We replace the heterogeneous knowledge network representation module in EMV based on the meta-path aggregation process with several SOTA heterogeneous graph networks. The experimental results are shown in Table 3. We find that the performance difference of each heterogeneous graph representation is insignificant, demonstrating that the heterogeneous knowledge representation based on the meta-path aggregation process could provide the most potent retrieval performance for EMV. In our analysis of the ablation results, a pronounced variation is discernible in the MR metric across different models and datasets. Such marked disparity in MR underscores its capability to provide a more granular retrieval performance evaluation. Indeed, this suggests that the MR metric serves as an efficacious discriminator, accentuating the nuanced differences in the performance dynamics across various models. Additionally, the impact of varying the number of layers (*i.e., k*) within the graph neural network on the efficacy of heterogeneous knowledge network representations remains nebulous. Consequently, empirical testing may be indispensable to pinpoint the optimal layer configuration for maximizing performance. Overall, for similar case retrieval tasks, the representation method based on the meta-path aggregation process enables faster aggregation of critical neighbor information through predefined meta-paths.

4.3.2. Sensitivity analysis

In practical scenarios, the model performance can be influenced by both the nuances of model parameters and the inherent differences within the data. We undertook a systematic investigation to elucidate these dynamics, focusing on pivotal factors that might influence similar case retrieval tasks. Specifically, we delve into the "Legal Document Length", "Embedding Dimension", and the "Number of Legal Entities" to discern their potential impacts. As depicted in Fig. 4, a comparative evaluation of model performances under various influencing factors is presented for a similar case retrieval task. Notably, EMV persistently outperforms the baseline models. Further, models enhanced with knowledge-based features manifest superior robustness and efficacy in their performance metrics. As the length of legal documents increases, the accuracy of all models shows a decreasing trend. For shorter legal documents, models predominantly reliant on text similarity, such as BERT, RoBERTa, and T5, exhibit superior retrieval precision. This might be attributed to the inherent simplicity of shorter documents, which may be adequately addressed through direct text comparison. However, when the length of legal documents exceeds 512, the performance of all models decreases significantly, with T5 decreasing by 24.59%, while KALM and EMV achieve performance reversal. This trend suggests that for longer legal documents, the capacity of knowledge-augmented models to extract latent entity and event information becomes pivotal,

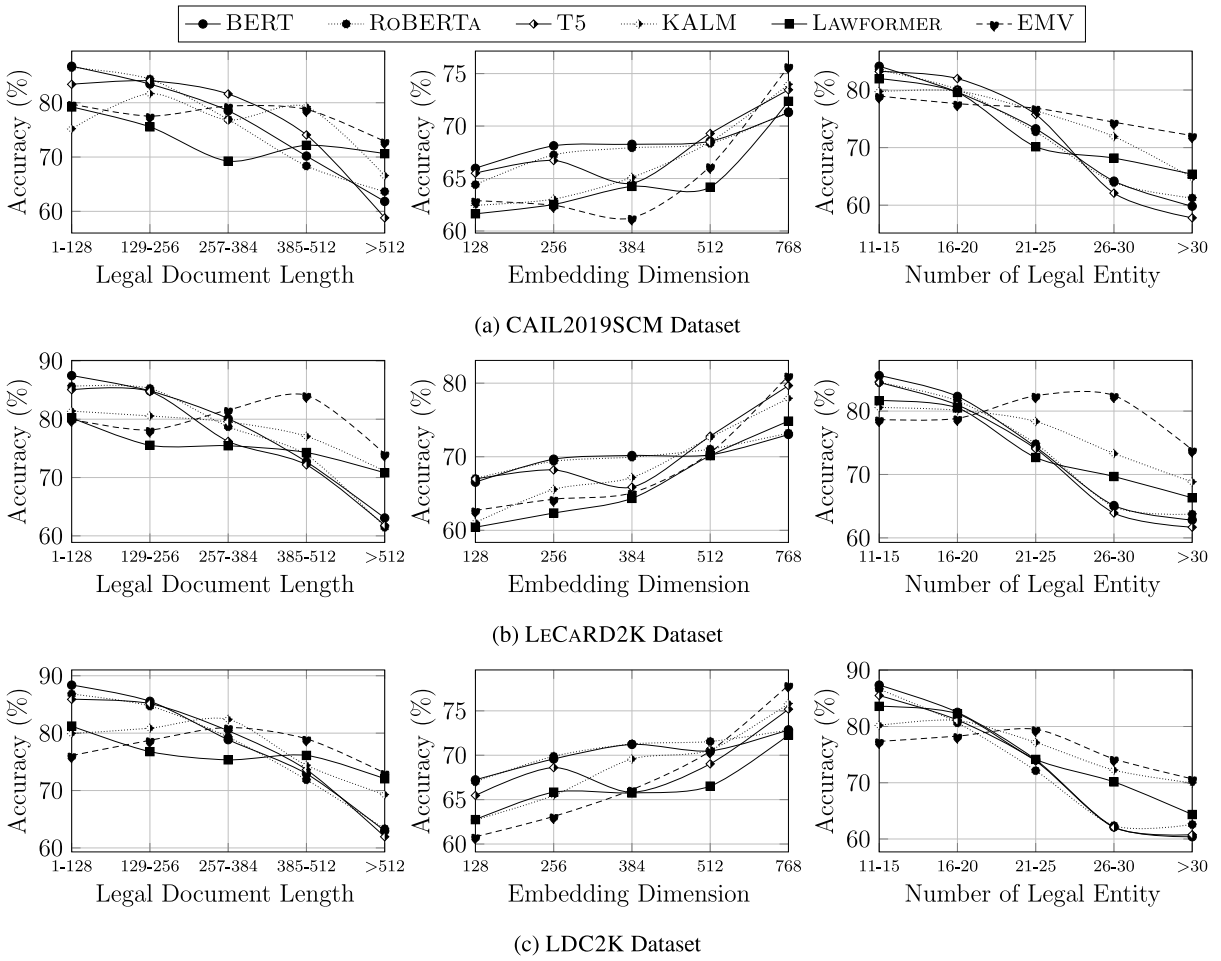
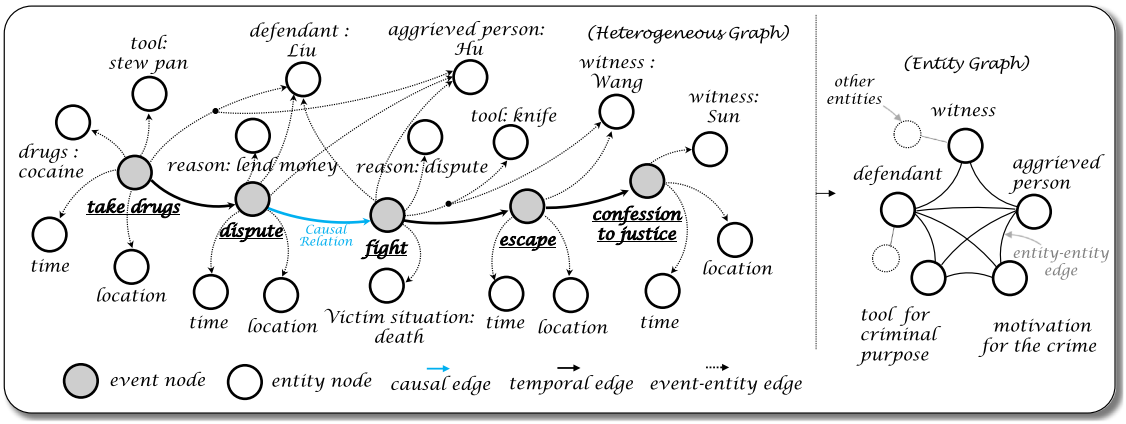


Fig. 4. Effects of "Legal Document Length", "Embedding Dimension" and "Number of Legal Entity" on the CAIL2019SCM, LeCaRD2K and LDC2K datasets.

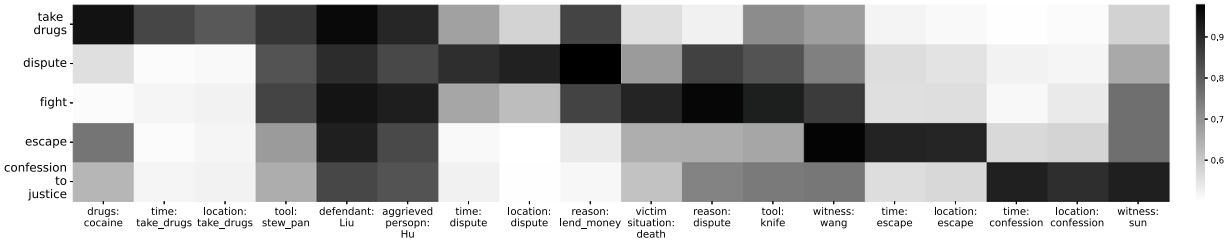
enabling them to facilitate case retrieval better. Upon escalating the embedding dimension, a pronounced upward trajectory in model accuracy becomes evident across all tested models, reaching a distinct zenith at the 768 dimension. Notably, KALM and EMV exhibit a heightened sensitivity to alterations in this dimension. This can likely be ascribed to their inherent architectural requisites since both need to introduce external knowledge and fuse feature representations from different semantic spaces. The experimental results show that keeping the vector dimensions as consistent as possible under different semantic spaces is beneficial to stimulating the performance of the models. For the number of legal entities within a document augments, a predominant trend of declining accuracy emerges across the majority of the evaluated models. Specifically, models *e.g.,* BERT, RoBERTa, T5 and Lawformer manifest heightened sensitivity to variations in entity count within legal documents, precipitating more rapid degradation in accuracy. Intriguingly, EMV stands in stark contrast, displaying robust resilience to such entity fluctuations, and notably clinching peak retrieval accuracy within the 21–25 and 26–30 entity brackets. Models such as KALM, architected to assimilate external entity-centric knowledge, can sometimes exhibit over-reliance or "paranoia" when inundated with excessive entities. This over-specification can be adroitly counterbalanced by the incorporation of event-centric knowledge.

4.3.3. Meta-path case analysis

The construction of the meta-paths is pivotal for the heterogeneous graph representation. In order to illustrate its effectiveness, we employ a heat map to delineate the attention weight distribution across varying neighborhood information within a heterogeneous graph instance. Notably, we adopt a meta-path *(EVE-ENT-ENT-ENT)* to capture the third-order entity neighborhood information of the event vertices, and retrain the model. As shown in Fig. 5, entity vertices {"drugs:cocaine", … , "witness:sun"} provide rich neighborhood information for event vertices {"Take drugs", … , "Confession to Justice"}. Through the heatmap, it can be observed that (1) even residing under the same meta path, the weights of different meta-path instances are different, such as ("Take drugs", "drugs: cocaine") and ("Take drugs", "location:take_drugs") in the meta-path π. This variation underlines the distinct



(a) Heterogeneous graph.



(b) Heatmap of the heterogeneous graph.

Fig. 5. Heatmap illustrating the attention weight distribution of event vertices with respect to entity vertices in a heterogeneous graph.

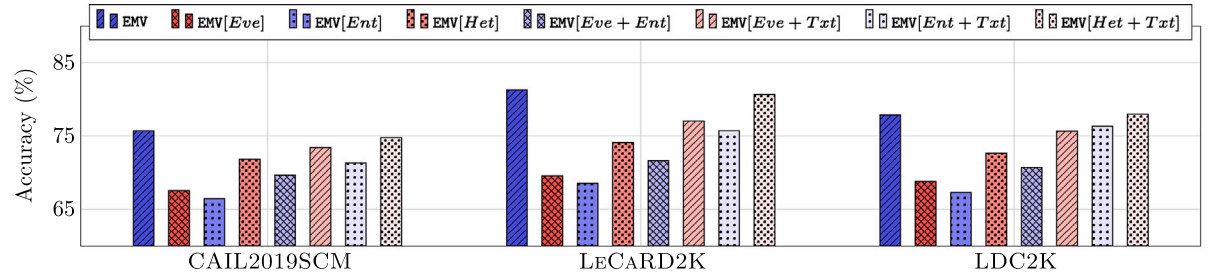
Table 4
Performance evaluation of various models on the SLCR task across different crime categories, namely, Larceny, Trafficking in Drugs, and Against Homeland Security. Bold values indicate the best performance in the respective category.

Model	Larceny		Trafficking in Drugs		Against Homeland Security	
	Acc.	MR	Acc.	MR	Acc.	MR
BERT	73.69	192.27	67.83	382.46	61.37	534.87
RoBERTa-base	73.98	187.36	68.33	361.38	61.46	463.46
T5-base	75.37	155.36	66.79	427.61	63.72	443.65
KALM	76.42	89.36	**74.84**	225.48	**68.26**	**223.29**
KnowBERT	71.24	152.36	70.26	249.33	65.73	416.10
LFESM	72.23	200.83	69.24	280.64	63.54	354.64
AlphaCourt	76.35	164.97	68.95	385.34	64.72	473.83
EMV	**83.60**	**58.86**	74.20	**163.59**	67.47	229.74

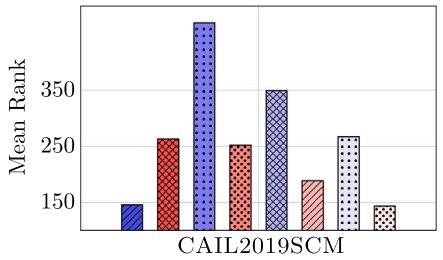
roles various neighboring vertices assume in event characterization, rendering a uniform attention weight distribution counterproductive. (2) The attention weight of second-order neighborhood information is noticeably subdued compared to its first-order counterpart, exemplified by contrasting the pairs (“Take drugs”, “drugs: cocaine”) and (“Take drugs”, “reason:lend_money”). Conversely, third-order neighborhood information is scarcely allocated any weight, as seen with pairs like (“Take drugs”, “location: escape”) and (“Take drugs”, “reason: dispute”), which shows that EMV has learned enough relevant entity information by encoding the first-order and second-order neighborhood. The “Time” and “Space” consumption can also be limited to an acceptable scope. (3) If there are multiple connective meta-paths between the event node and the target entity node, then the event node’s attention weight for that entity node is significantly enhanced, such as (“Take drugs”, “reason:lend_money”), (“Take drugs”, “tool: knife”). For the two meta-paths that aggregate second-order neighborhood information, the meta-path τ exerts a more pronounced influence on event information representation. In addition, we have also verified through experiments that considering the third-order neighborhood information and using the average instead of the attention mechanism to calculate the weight both damage the performance of EMV, which also verifies the rationality of the meta-paths we designed.

4.3.4. Crime analysis

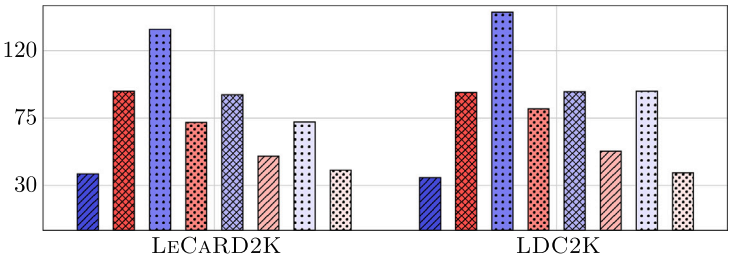
Legal documents containing the same type of crime may not be similar cases because of differences in the details of the case. In the context of similar legal case retrieval tasks, it is evident that distinct crimes possess inherently differentiated case information,



(a) Assessment Outcomes for the Accuracy of Individual Modules.



(b) Mean Rank metric of Individual Modules.



(c) Mean Rank metric of Individual Modules.

Fig. 6. The assessment outcomes derived from ablation experiments for individual component modules of the EMV model in diverse datasets.

Table 5
Performance of the EMV model under variations of the training dataset: examining the impact of random removal of 20% of entity or event nodes and the systematic exclusion of 20% of high-frequency entity nodes on the similar case retrieval task in diverse datasets.

Model	CAIL2019SCM		LeCaRD2K		LDC2K	
	Acc.	MR	Acc.	MR	Acc.	MR
EMV	**75.70** (±0.6)	145.55 (±7.3)	**81.30** (±0.8)	**37.85** (±2.8)	**77.90** (±0.3)	35.27 (±3.6)
EMV[*Ent_rand*20%]	$73.12\downarrow_{2.58\%}$	$164.26\downarrow_{11.39\%}$	$78.25\downarrow_{3.05\%}$	$64.58\downarrow_{41.39\%}$	$75.85\downarrow_{2.05\%}$	$72.14\downarrow_{51.11\%}$
EMV[*Ent_hf*20%]	$75.49\downarrow_{0.21\%}$	$126.04\uparrow_{15.48\%}$	$79.53\downarrow_{1.77\%}$	$44.31\downarrow_{14.58\%}$	$77.27\downarrow_{0.63\%}$	$32.74\uparrow_{7.73\%}$
EMV[*Eve_rand*20%]	$71.86\downarrow_{3.84\%}$	$191.65\downarrow_{24.05\%}$	$74.17\downarrow_{7.13\%}$	$82.15\downarrow_{53.93\%}$	$74.09\downarrow_{3.81\%}$	$81.27\downarrow_{56.60\%}$

rendering the assessment of case similarity relatively straightforward. When delving into cases pertaining to the same crime type, the nuances in core circumstances or evidentiary data often converge, obfuscating clear distinctions in their similarities. Moreover, in the process of curating our datasets, we discern pronounced distributional disparities amongst the legal documents containing different types of crimes. For instance, legal documents containing larceny-related offenses constituted approximately 30% of the overall dataset, and those containing against homeland security were markedly sparse, representing a mere 3.2%. Such a stark imbalance in data distribution possesses the potential to introduce significant biases, thereby influencing the integrity of experimental outcomes.

To further clarify, we hone in on three emblematic crimes for our analysis, including larceny, drug trafficking, and against homeland security. Experimental data corresponding to these crimes were meticulously curated, delineated in Table 4. For larceny crime, the EMV model distinctly outperformed the baseline models. This superior performance can be attributed to the event information's pivotal role in streamlining the voluminous retrieval of candidate documents pertinent to high-frequency crimes. For drug-related crime, a marked performance chasm is evident between the EMV and other baseline models. This can be ascribed to the event information's capability to counteract potential misconceptions induced by high-frequency entity information. With respect to crimes against homeland security, the overarching performance metrics of the baseline models tend to dip in comparison to other crimes. This trend is primarily rooted in the dearth of adequately labeled corpora. Nevertheless, even amidst these constraints, EMV demonstrates a commendable edge. The model's integration of multi-level knowledge information proved instrumental in mitigating challenges posed by the dearth of reference benchmarks for low-frequency crimes.

4.4. Ablation analysis

Fig. 6 presents an in-depth ablation analysis results of the EMV model, illustrating its performance across various datasets. By systematically removing different components of the model, we can discern the contributions of each component to the overall

performance. Let EMV[L] represent a specific variant of the EMV model, where L belongs to the set {*Txt,Ent,Eve,Het*}. Here, *Txt*, *Ent*, *Eve*, and *Het* stand for text, the entity knowledge network, the event knowledge network, and the heterogeneous knowledge network modules, respectively.

From Fig. 6, we observe that any sub-module of EMV can help the matching results. The complete EMV model showcases superior accuracy, especially on the CAIL2019SCM and LeCaRD2K datasets, indicating that the integration of all features substantially contributes to its high performance. When either event information (*Eve*) or entity information (*Ent*) is solely utilized, there is a noticeable drop in performance across datasets compared to the complete model. This indicates that both event and entity representations are crucial for the model to perform optimally. Utilizing heterogeneous information (*Het*) alone results in a noticeable performance increase compared to using just event or entity information, proving that entity knowledge and event knowledge belong to different levels of semantic information and that there is an implicit semantic correlation between them. EMV[*Het+Txt*], EMV[*Ent+Txt*], and EMV[*Eve+Txt*] have apparent performance improvement, which shows that both entity knowledge and event knowledge can help the model obtain richer semantic features.

We devise an interesting experiment to delve deeper into the influence of entity and event knowledge embedded within legal documents on a similar case retrieval task. Specifically, we subject the training dataset to alterations by randomly eliminating 20% of either the entity or event nodes. Further augmenting our analysis, we conducted an additional experiment wherein 20% of the high-frequency entity nodes were systematically removed from the training set. The repercussions of these modified training datasets on the EMV model's performance are detailed in Table 5. The results presented in Table 5 show that EMV achieves its peak performance when trained on the unaltered training set. This underscores the value of incorporating both entity and event information, highlighting their synergistic effect in enhancing the retrieval capabilities of the EMV model. Interestingly, compared with randomly deleting entity nodes, the deletion of 20% high-frequency entity nodes results in a negligible model performance degradation. Especially on the CAIL2019SCM and LDC2K datasets, the MR metrics even witnessed considerable improvement. This indicates that high-frequency entities may carry redundant noise that biases the model. Our analysis reveals that eliminating a mere 20% of event nodes introduces a pronounced detrimental effect on the performance of EMV, indicating the indispensable role that event nodes have in this model.

5. Conclusion

In this paper, we primarily focused on resolving the task of Similar Legal Case Retrieval (SLCR) to improve the identification of relevant judicial precedents and enhance the trial credibility and fairness. Although this task is challenging due to the intricateness and multifacetedness of the whole legal case structure, the "fact description" part in legal documents encapsulates the essence of cases and serves as the basis for measuring similarity. To comprehensively depict this critical part, we focused on exploring the event knowledge contained within this part and represented it by the event graph, where the events and actions within legal cases were treated as nodes and edges. Additionally, the heterogeneous knowledge graph was further employed to incorporate event knowledge, which could effectively encode the latent semantic features between entities and events. By integrating these two types of exploration in event knowledge, a novel model was proposed, dubbed EMV (*i.e.,* the acronym of *Event is More Valuable*), enabling multiple perspectives on legal cases and improving the model's ability to retrieve the optimal similar cases. Experimental results on multiple similar legal case retrieval datasets have demonstrated the effectiveness of EMV, since it could surpass the state-of-the-art model by a substantial margin of 4.3% in the SLCR task.

Reproducibility

Our datasets and codes include: (1) **CAIL2019SCM** dataset, **LeCaRD2K** dataset and **LDC2K** dataset; (2) Codes for EMV model; (3) Codes used to evaluate MR indicators; (4) TransE embedding vectors for the knowledge graph LegalKG. The above datasets and codes are released at https://github.com/YuxinZhangGit/EMV.git.

CRediT authorship contribution statement

Yuxin Zhang: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Songlin Zhai:** Data curation, Writing – review & editing. **Yuan Meng:** Investigation. **Sheng Bi:** Validation. **Yongrui Chen:** Validation. **Guilin Qi:** Supervision, Validation, Writing – review & editing.

Data availability

The datasets and codes are released at the link: https://github.com/YuxinZhangGit/EMV.

Acknowledgments

This work is partially supported by the National Nature Science Foundation of China under No. U21A20488. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Alec, R., Jeffrey, W., Rewon, C., David, L., Dario, A., & Ilya, S. (2019). Language models are unsupervised multitask learners. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Bench-Capon, T. J. M., Araszkiewicz, M., Ashley, K. D., Atkinson, K., Bex, F., Borges, F., et al. (2012). A history of AI and law in 50 papers: 25 years of the international conference on AI and law. *Artificial Intelligence and Law*, *20*(3), 215–319. http://dx.doi.org/10.1007/S10506-012-9131-X.

Bhattacharya, P., Ghosh, K., Pal, A., & Ghosh, S. (2020). Hier-spcnet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2020, virtual event, China, July 25-30, 2020* (pp. 1657–1660). ACM, http://dx.doi.org/10.1145/3397271.3401191.

Bhattacharya, P., Ghosh, K., Pal, A., & Ghosh, S. (2022). Legal case document similarity: You need both network and text. *Information Processing & Management*, *59*(6), Article 103069. http://dx.doi.org/10.1016/J.IPM.2022.103069.

Bi, S., Ali, Z., Wang, M., Wu, T., & Qi, G. (2022). Learning heterogeneous graph embedding for Chinese legal document similarity. *Knowledge-Based Systems*, *250*, Article 109046. http://dx.doi.org/10.1016/j.knosys.2022.109046.

Bi, S., Huang, Y., Cheng, X., Wang, M., & Qi, G. (2019). Building Chinese legal hybrid knowledge network. In C. Douligeris, D. Karagiannis, & D. Apostolou (Eds.), *Lecture notes in computer science*: *vol. 11775*, *Knowledge science, engineering and management - 12th international conference, KSEM 2019, athens, Greece, August 28-30, 2019, proceedings, part i* (pp. 628–639). Springer, http://dx.doi.org/10.1007/978-3-030-29551-6_56.

Cai, Z., He, Z., Guan, X., & Li, Y. (2018). Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 577–590. http://dx.doi.org/10.1109/TDSC.2016.2613521.

Cai, Z., & Zheng, X. (2020). A private and efficient mechanism for data uploading in smart cyber-physical systems. *IEEE Transactions on Network Science and Engineering*, *7*(2), 766–775. http://dx.doi.org/10.1109/TNSE.2018.2830307.

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), Article 102798. http://dx.doi.org/10.1016/J.IPM.2021.102798.

Devlin, J., Chang, M., 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Feng, S., Tan, Z., Zhang, W., Lei, Z., & Tsvetkov, Y. (2023). KALM: knowledge-aware integration of local, document, and global contexts for long document understanding. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), ACL 2023, Toronto, Canada, July 9-14, 2023* (pp. 2116–2138). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.118.

Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., et al. (2023). What is event knowledge graph: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 7569–7589. http://dx.doi.org/10.1109/TKDE.2022.3180362.

Hong, Z., Zhou, Q., Zhang, R., Li, W., & Mo, T. (2020). Legal feature enhanced semantic matching network for similar case matching. In *2020 international joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/IJCNN48605.2020.9207528.

Hu, W., Zhao, S., Zhao, Q., Sun, H., Hu, X., Guo, R., et al. (2022). BERT_LF: A similar case retrieval method based on legal facts. *Wireless Communications and Mobile Computing*, *2022*, http://dx.doi.org/10.1155/2022/2511147.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a meeting of SIGDAT, a special interest group of the ACL* (pp. 1746–1751). ACL, http://dx.doi.org/10.3115/v1/d14-1181.

Kumar, S., Reddy, P. K., Reddy, V. B., & Suri, M. (2013). Finding similar legal judgements under common law system. In A. Madaan, S. Kikuchi, & S. Bhalla (Eds.), *Lecture notes in computer science*: *vol. 7813*, *Databases in networked information systems - 8th international workshop, DNIS 2013, Aizu-Wakamatsu, Japan, March 25-27, 2013. proceedings* (pp. 103–116). Springer, http://dx.doi.org/10.1007/978-3-642-37134-9_9.

Li, H., Lu, J., Le, Y., & He, J. (2022). IACN: interactive attention capsule network for similar case matching. *Intelligent Data Analysis*, *26*(2), 525–541. http://dx.doi.org/10.3233/IDA-205632.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR arXiv:1907.11692, URL http://arxiv.org/abs/1907.11692.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In S. Kambhampati (Ed.), *Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016* (pp. 2873–2879). IJCAI/AAAI Press, URL http://www.ijcai.org/Abstract/16/408.

Liu, X., Wu, K., Liu, B., & Qian, R. (2023). HNERec: Scientific collaborator recommendation model based on heterogeneous network embedding. *Information Processing & Management*, *60*(2), Article 103253. http://dx.doi.org/10.1016/J.IPM.2022.103253.

Lou, D., Liao, Z., Deng, S., Zhang, N., & Chen, H. (2021). Mlbinet: A cross-sentence collective event detection network. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL/IJCNLP 2021, (volume 1: long papers), virtual event, August 1-6, 2021* (pp. 4829–4839). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2021.ACL-LONG.373.

Lu, Q., & Conrad, J. G. (2012). Bringing order to legal documents - an issue-based recommendation system via cluster association. In J. Filipe, & J. L. G. Dietz (Eds.), *KEOD 2012 - proceedings of the international conference on knowledge engineering and ontology development, Barcelona, Spain, 4 - 7 October, 2012* (pp. 76–88). SciTePress.

Lyu, Y., Wang, Z., Ren, Z., Ren, P., Chen, Z., Liu, X., et al. (2022). Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing & Management*, *59*(1), Article 102780. http://dx.doi.org/10.1016/J.IPM.2021.102780.

Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M., et al. (2021). Lecard: A legal case retrieval dataset for Chinese law system. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), *SIGIR '21: the 44th international ACM SIGIR conference on research and development in information retrieval, virtual event, Canada, July 11-15, 2021* (pp. 2342–2348). ACM, http://dx.doi.org/10.1145/3404835.3463250.

Peters, M. E., Neumann, M., IV, R. L. L., Schwartz, R., Joshi, V., Singh, S., et al. (2019). Knowledge enhanced contextual word representations. In K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (pp. 43–54). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1005.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 140:1–140:67, URL http://jmlr.org/papers/v21/20-074.html.

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., et al. (2020). BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020* (pp. 3501–3507). ijcai.org, http://dx.doi.org/10.24963/IJCAI.2020/484.

Shen, S., Qi, G., Li, Z., Bi, S., & Wang, L. (2020). Hierarchical Chinese legal event extraction via pedal attention mechanism. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics, COLING 2020, Barcelona, Spain (online), December 8-13, 2020* (pp. 100–113). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.9.

Shen, S., Zhou, H., Wu, T., & Qi, G. (2022). Event causality identification via derivative prompt joint learning. In N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022* (pp. 2288–2299). International Committee on Computational Linguistics, URL https://aclanthology.org/2022.coling-1.200.

Song, X., Li, J., Cai, T., Yang, S., Yang, T., & Liu, C. (2022). A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, *258*, Article 110036. http://dx.doi.org/10.1016/J.KNOSYS.2022.110036.

Sun, Z., Huang, J., Xu, X., Chen, Q., Ren, W., & Hu, W. (2023). What makes entities similar? A similarity flooding perspective for multi-sourced knowledge graph embeddings. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of machine learning research*: *vol. 202*, *International conference on machine learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA* (pp. 32875–32885). PMLR, URL https://proceedings.mlr.press/v202/sun23d.html.

Trivedi, A., Trivedi, A., Varshney, S., Joshipura, V., Mehta, R., & Dhanani, J. (2021). Similarity analysis of legal documents: A survey. In *ICT analysis and applications* (pp. 497–506). Springer, http://dx.doi.org/10.1007/978-981-15-8354-449.

van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, *25*(1), 65–87. http://dx.doi.org/10.1007/S10506-017-9195-8.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., et al. (2019). Heterogeneous graph attention network. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), *The world wide web conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (pp. 2022–2032). ACM, http://dx.doi.org/10.1145/3308558.3313562.

Winkels, R., Boer, A., Vredebregt, B., & van Someren, A. (2014). Towards a legal recommender system. In R. Hoekstra (Ed.), *Frontiers in artificial intelligence and applications*: *271*, *Legal knowledge and information systems - JURIX 2014: the twenty-seventh annual conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014* (pp. 169–178). IOS Press, http://dx.doi.org/10.3233/978-1-61499-468-8-169.

Wu, Z., Mao, J., Liu, Y., Zhan, J., Zheng, Y., Zhang, M., et al. (2020). Leveraging passage-level cumulative gain for document ranking. In *Proceedings of the web conference 2020* (pp. 2421–2431). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3366423.3380305.

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, *2*, 79–84. http://dx.doi.org/10.1016/j.aiopen.2021.06.003.

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., et al. (2019). CAIL2019-SCM: a dataset of similar case matching in legal domain. CoRR arXiv:1911.08962, URL http://arxiv.org/abs/1911.08962.

Yun, S., Jeong, M., Kim, R., Kang, J., & Kim, H. J. (2019). Graph transformer networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, neurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 11960–11970). URL https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html.

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, & G. Karypis (Eds.), *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019* (pp. 793–803). ACM, http://dx.doi.org/10.1145/3292500.3330961.

Zhang, H., Zhang, Z., Liu, Z., & Sun, M. (2019). Open Chinese language pre-trained model zoo. *Technical Report*, Natural Language Processing and Social Humanistic Computing Center, Tsinghua University, URL https://github.com/thunlp/openclap.

Zhao, Q., Gao, T., & Guo, N. (2023). LA-MGFM: a legal judgment prediction method via sememe-enhanced graph neural networks and multi-graph fusion mechanism. *Information Processing & Management*, *60*(5), Article 103455. http://dx.doi.org/10.1016/J.IPM.2023.103455.