Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9064

Similar-Case Retrieval for Government Hotlines via an Appeal-Driven Event Graph and Dual-Route Hybrid Retrieval

Yizhou Fang*. Rencheng Sun
Qingdao University of Computer Science and Technology, QingDao 266071, China
Email: 1905288354@qq.com

How to cite this paper: Fang, Y. Z., & Sun, R. C. (2026). Similar-Case Retrieval for Government Hotlines via an Appeal-Driven Event Graph and Dual-Route Hybrid Retrieval. *Journal of Computer Science and Frontier Technologies*, 3(1), 112-126. ISSN Print: 3104-4204; ISSN Online: 3104-4212. https://doi.org/10.63313/JCSFT.9064
Published: 2026-04-24

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

To address the colloquial expressions, redundant descriptions, and dispersed structural information commonly found in historical work orders from 12345 government hotlines, traditional keyword matching and purely semantic retrieval methods often fail to identify cases that are both semantically relevant and structurally comparable, limiting the effectiveness of similar-case retrieval in complex hotline scenarios. This paper proposes a similar-case retrieval method for government hotlines based on an appeal-driven event graph and dual-route hybrid retrieval. Specifically, key elements, including location, time, core issue, and core appeal, are extracted from historical work orders to construct an appeal-driven complaint event knowledge graph. On this basis, a dual-route hybrid retrieval mechanism that integrates dense vector recall and graph-based logical recall is designed, and a structure-aware fusion strategy with cross-encoder reranking is further introduced to improve candidate ranking quality. Experiments on a real government hotline complaint dataset and the public LeCaRD legal case retrieval dataset show that the proposed method outperforms multiple baseline methods. On the complaint dataset, it achieves a P@5 of 0.613 and a Mean Average Precision(MAP) of 0.633, while on LeCaRD it reaches a P@5 of 0.435 and a MAP of 0.487. The results demonstrate that the proposed method can effectively improve the accuracy and structural comparability of similar-case retrieval in complex government hotline scenarios.

Keywords

Government Hotline; Similar-Case Retrieval; Appeal-Driven Event Graph; Dual-Route Hybrid Retrieval; Structure-Aware Reranking

1. Introduction

With the continuous advancement of digital government and intelligent grassroots governance, China's 12345 government service hotline has become an important channel for governments to perceive social needs, respond to public appeals, and

promote collaborative governance[1][2]. Hotline work orders involve a wide range of domains, including urban management, public services, and market supervision, and are characterized by diverse sources, complex categories, and rapid updates[3].As the volume of hotline work orders continues to grow, retrieving historically comparable cases from large-scale case repositories has become increasingly important for improving the efficiency and intelligence of hotline handling. High-quality similar-case retrieval can provide valuable support for complaint understanding, case assessment, experience reuse, and subsequent decision-making.

Different from general text retrieval tasks, government hotline work orders are usually described directly by citizens in natural language. As a result, the texts often contain colloquial expressions, emotional statements, redundant descriptions, and fragmented factual information. Even for the same type of issue, different callers may use very different expressions. In practice, however, whether two cases are truly comparable depends not only on superficial lexical similarity, but also on the consistency of key event elements, such as the core issue, core appeal, and spatiotemporal conditions. If retrieval relies only on surface-level textual features or overall semantic similarity, it may return cases that appear topically similar but are not structurally comparable, thereby weakening the practical value of historical cases.

Existing text retrieval methods have generally evolved from lexical matching to semantic modeling and then to hybrid retrieval. Early methods such as Term Frequency-Inverse Document Frequency(TF-IDF) and Okapi BM25(BM25) are efficient and easy to implement, but they rely heavily on literal word overlap and thus struggle with non-standard expressions in hotline complaints[4][5]. With the development of pretrained language models such as Bidirectional Encoder Representations from Transformers (BERT) and Sentence-BERT, dense retrieval methods have shown stronger capability in matching semantically similar texts with different surface expressions[6][7]. Nevertheless, in the government hotline scenario, semantic similarity alone is still insufficient. On the one hand, it cannot explicitly capture structured case elements such as location, time, core issue, and appeal. On the other hand, it is relatively weak in modeling the consistency of event composition and case logic. Therefore, an effective retrieval method should not only preserve semantic coverage, but also incorporate stronger structural representation and logical relevance.

Knowledge graphs provide a promising way to organize complex case information in a structured form[8]. By explicitly representing key case elements, such as location, time, core issue, and core appeal, as nodes and relations, the dispersed factual information embedded in natural language descriptions can be transformed into a computable graph structure. On this basis, combining dense semantic recall with graph-based logical recall makes it possible to jointly model similarity at the

expression level and comparability at the event-structure level. Furthermore, reranking mechanisms can be introduced to refine the final ranking of retrieved cases and improve retrieval precision[9].

Motivated by these observations, this paper proposes a similar-case retrieval method for government hotlines based on an appeal-driven event graph and dual-route hybrid retrieval. Specifically, key elements are extracted from historical work orders to construct an appeal-driven complaint event knowledge graph. A dual-route hybrid retrieval mechanism that integrates dense vector recall and graph-based logical recall is then designed, followed by structure-aware fusion and cross-encoder reranking to improve the overall quality of candidate ranking. Experiments conducted on a real government hotline complaint dataset and the public LeCaRD legal case retrieval dataset[10] demonstrate that the proposed method consistently outperforms multiple baseline methods, showing its effectiveness in improving the accuracy and structural comparability of similar-case retrieval in complex government hotline scenarios.

2. Problem Analysis

Similar-case retrieval in 12345 government hotline scenarios is not a conventional text similarity matching task. Citizens usually describe their complaints in natural language, and hotline work orders often contain a mixture of problem background, emotional expressions, expected handling outcomes, and spatiotemporal information. As a result, the texts are typically colloquial, redundant, and fragmented. Different callers may describe the same issue in very different ways, while the actual value of a historical case does not primarily depend on lexical resemblance, but on whether the two cases are comparable in terms of their core issue, major appeal, and situational context. If these key facts cannot be accurately identified from complex descriptions, retrieval results are likely to remain at the level of superficial topical similarity and may provide limited support for case referencing and knowledge reuse.

From the perspective of existing retrieval methods, traditional lexical approaches such as TF-IDF and BM25 are efficient and easy to implement, but they rely heavily on surface-level word matching. Consequently, they are prone to missing relevant cases when faced with non-standard expressions, synonymous reformulations, and colloquial descriptions in public complaints. With the development of pretrained language models, dense semantic retrieval methods based on BERT or sentence embeddings have achieved better performance in cross-expression matching and can alleviate the problem of lexical inconsistency to a certain extent. However, these methods still mainly operate on overall semantic similarity. Although they can identify historically similar cases at the semantic level, they cannot guarantee consistency in event composition and case context. In government hotline scenarios, semantic similarity alone does not necessarily imply case comparability. Without an

explicit representation of internal case structure, retrieval systems may still return cases that are topically similar but factually inconsistent.

Another important challenge is that government hotline cases are highly context-dependent. Whether a historical work order can serve as a useful reference depends not only on the topic described in the text, but also on event elements such as location, time, core issue, and core appeal. In other words, similar-case retrieval should not merely answer the question of which texts look similar; it should further identify which cases are structurally closer and therefore more useful as references. If a retrieval model cannot explicitly characterize the relationships among these key elements, even semantically relevant results may fail to provide practical support because of differences in appeal objectives, issue focus, or spatiotemporal conditions. Therefore, improving the structural comparability between cases is a central issue in similar-case retrieval for complex government hotline scenarios.

Based on the above analysis, this study argues that similar-case retrieval for government hotlines mainly faces two core problems. The first is how to extract key elements, such as location, time, core issue, and core appeal, from unstructured work-order texts and organize them into a computable and associated structured representation. The second is how to incorporate structural consistency into the retrieval process while preserving semantic coverage, so that the returned results can both cover semantically similar cases with different expressions and identify historical cases that are truly close in event composition and appeal objectives. Accordingly, this paper formulates government hotline similar-case retrieval as a structured similar-case ranking task for complex case texts. Given a query work order, the goal is to retrieve the Top-K historical cases that are as consistent as possible with the query in terms of core issue, appeal objective, and spatiotemporal conditions, and to build the subsequent method design on this basis.

3. Method

3.1. Overall Framework

This study proposes a dual-route hybrid retrieval method based on an appeal-driven event knowledge graph for similar-case retrieval in complex 12345 government hotline scenarios. Unlike conventional approaches that rely solely on keyword matching or a single semantic retrieval pathway, the proposed method formulates similar-case retrieval as a unified process consisting of event element extraction, graph-based case representation, dual-route collaborative recall, and structure-aware fusion reranking. Its main objective is to retrieve historical cases that are not only semantically similar to the query work order, but also highly comparable in terms of event structure.

Specifically, the proposed method contains three major components. First, at the level of historical case representation, key event elements, including location, time, core issue, and core appeal, are extracted from work-order texts. These extracted

elements are then combined with basic business fields, such as case title, category hierarchy, and historical handling department, to construct an appeal-driven event knowledge graph centered on individual cases. In this way, the originally dispersed information embedded in natural language descriptions is transformed into a structured and associated representation, enabling case similarity to be characterized not only by overall textual similarity but also by shared issues, shared appeals, and shared spatiotemporal conditions.

Second, at the candidate recall stage, a dual-route retrieval mechanism is designed by combining semantic recall and graph-based logical recall. The semantic recall branch is used to quickly identify candidate cases that are close to the query at the expression level, thereby ensuring sufficient semantic coverage. The graph-based logical recall branch further supplements the candidate set by exploring structural relations in the event graph, allowing the method to retrieve cases that may not be highly similar in surface wording but are much closer in issue composition, appeal objective, and spatiotemporal context. The two branches correspond to semantic nearest-neighbor search in text space and structural complementary search in graph space, respectively, and jointly improve the quality of similar-case retrieval.

Finally, at the ranking stage, because the candidates returned by the semantic branch and the graph branch emphasize different aspects of similarity, a structure-aware fusion and cross-encoder reranking mechanism is introduced to generate the final ranking. By jointly considering semantic relevance, graph-based relevance, and interaction-based reranking scores, the method corrects ranking errors caused by cases that are semantically similar but factually inconsistent, or structurally related but lexically dissimilar. As a result, the final Top-K retrieved cases are both semantically relevant and structurally comparable, which provides stronger support for case referencing and experience reuse in government hotline scenarios. The overall framework of the proposed method is illustrated in Figure 1.

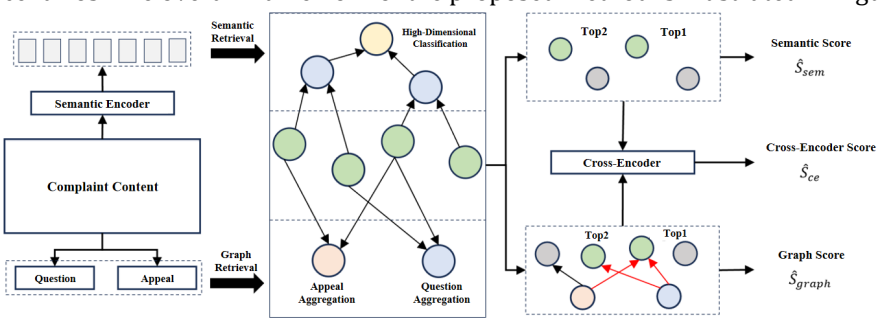


Figure 1. Dual-route Hybrid Retrieval Framework

3.2. Dual-Route Hybrid Retrieval Based on the Event Graph

3.2.1. Construction of the Appeal-Driven Event Knowledge Graph

To improve the structural comparability of complex work orders, this study first constructs an appeal-driven event knowledge graph for similar-case retrieval.

Specifically, each historical work order is treated as an individual case node, while basic business fields, including case title, category hierarchy, and historical handling department, are preserved. On this basis, key event elements, such as location, time, core issue, and core appeal, are further extracted from the main complaint text and organized as multi-type attribute nodes and relation edges. In this way, the originally dispersed information in natural language descriptions is transformed into a graph structure of "case–issue–appeal–spatiotemporal condition", enabling case relations to be explicitly represented through shared issues, shared appeals, and shared spatiotemporal elements rather than relying only on overall text similarity.

At the stage of basic graph construction, three types of business information are retained. The first is the category hierarchy, which reflects the issue domain of a case. The second is the historical handling department, which preserves the organizational attribute of the case in the business system. The third is the work-order title, which provides a concise summary of the case topic and assists subsequent retrieval. Together, these fields and the case node form the basic entity layer of the graph and provide the basis for later event-element integration and graph-based retrieval. To ensure consistency between graph storage and graph computation, a unified representation is adopted for graph relations, and edges are not forcibly created when a field is missing, thereby reducing pseudo-connections and the risk of noise propagation in subsequent graph retrieval.

At the event-element extraction stage, a large language model is used to conduct structured parsing of the main complaint text. The extracted information mainly includes four key semantic slots: location, time, core issue, and core appeal. Location represents the territorial scope and governance unit involved in the case, time characterizes the temporal context of the event, core issue summarizes the central conflict reflected in the complaint, and core appeal indicates the handling action expected by the citizen. These four types of information together constitute the most discriminative event-level semantic structure of a case and provide important support for subsequent graph-based retrieval.

Considering that government hotline texts often contain colloquial expressions, abbreviations, non-standard wording, and missing fields, strict string matching is not adopted during graph construction. Instead, semantic normalization and similarity alignment are applied to nodes such as location, issue, and appeal. For example, street aliases and abbreviated residential-area names are linked according to semantic similarity rather than exact lexical identity. In addition, the extracted structured results are constrained through a unified JSON format. If a certain field cannot be reliably identified from the text, it is recorded as None or no corresponding relation is generated. These strategies reduce graph fragmentation caused by non-standard expressions and improve the adaptability of subsequent retrieval to real-world hotline texts.

After the basic entity relations and event elements are constructed, they are integrated into a complete appeal-driven event knowledge graph. The key idea of this integration is to take the case node as the center and jointly connect business-field relations, such as category hierarchy, department, and title, with event-semantic relations extracted by the large language model, including time, location, core issue, and appeal, into a unified graph structure. As a result, each case node contains both its structural attributes in the business system and its event-semantic attributes in the complaint text, thereby providing a unified knowledge representation for subsequent dual-route hybrid retrieval.

3.2.2. Collaborative Semantic Recall and Graph-Based Logical Recall

After constructing the case graph, a dual-route hybrid retrieval mechanism is designed to collaboratively recall candidate cases from both semantic and structural perspectives. The semantic recall branch is introduced to handle the diversity of citizen expressions, frequent paraphrases, and inconsistent wording. The graph-based logical recall branch is intended to compensate for the limitations of purely semantic retrieval in modeling event structure and logical consistency. In other words, semantic recall emphasizes coverage, while graph-based logical recall emphasizes structural comparability. Together, they determine the quality of the final candidate set.

For a query work order q and a candidate work order d, both are first encoded by a sentence encoder $E(\cdot)$ into dense vector representations:

$$x_q = E(q), x_d = E(d)$$

Then, semantic relevance is calculated using cosine similarity:

$$S_{sem}(x_{\mathrm{q}}, x_{\mathrm{d}}) = \frac{{x_{\mathrm{q}}}^T x_{\mathrm{d}}}{||x_{\mathrm{q}}||\,||x_{\mathrm{d}}||}$$

Candidate cases are ranked according to the semantic score to obtain the semantic candidate set. This branch is able to cover historical cases that are expressed differently but remain semantically close to the query, and therefore serves as the basic recall path of the whole framework. At the same time, it is computationally efficient and suitable for large-scale initial recall over historical case repositories.

On this basis, graph-based logical recall is further performed on the event knowledge graph. Specifically, key elements, such as location, time, core issue, and core appeal, are first identified from the query work order, and corresponding seed nodes are matched in the graph. Starting from these seed nodes, the method searches their connected case nodes and related attribute nodes in local subgraphs, so as to capture structural co-occurrence relations between the query case and historical cases. For a candidate case d, the graph-based relevance score is defined as

$$S_{graph}(q,d) = \alpha\, S_{dir}(q,d) + (1-\alpha) S_{prop}(q,d)$$

where α is a balancing coefficient,$S_{dir}(q,d)$ denotes the direct matching strength between the query elements and the candidate case in the graph, and $S_{prop}(q,d)$

denotes the indirect structural evidence obtained by local subgraph propagation. The propagation score is computed as

$$S_{prop}(q,d) = (P^n h_0)_d$$

where h_0 is the initial weight vector of seed nodes,P is the normalized transition matrix of the graph, and n is the propagation step. Unlike directly counting matched attributes, the propagation score incorporates indirect structural information induced by historical case co-occurrence relations. Therefore, this branch can identify cases that are not highly similar in wording but are strongly aligned in issue composition, appeal objectives, and spatiotemporal conditions, thereby improving structural comparability in similar-case retrieval.

3.2.3. Structure-Aware Fusion and Reranking

Because the candidate cases returned by the semantic branch and the graph branch focus on different aspects of similarity, further unified fusion and fine-grained reranking are required to improve the final ranking quality. After dual-route recall, the semantic candidate set and the graph-based candidate set are merged into a unified candidate pool. A cross-encoder reranking model is then introduced to perform fine-grained interaction modeling between the query work order and each candidate case, producing an interaction-based reranking score. Finally, a unified ranking function is constructed by combining semantic relevance, graph relevance, and reranking evidence.The final ranking score is defined as

$$s_{final} = \lambda_{ce}\hat{s}_{ce} + \lambda_{sem}\hat{s}_{sem} + \lambda_{graph}\hat{s}_{graph}$$

where $\hat{s}_{ce}$,$\hat{s}_{sem}$ and $\hat{s}_{graph}$ represent the normalized cross-encoder score, semantic recall score, and graph-based recall score, respectively, and λ_{ce},λ_{sem} and λ_{graph} are tunable weights. Under this fusion mechanism, reranking plays the leading role, while semantic evidence and structural evidence serve as stable complementary signals. As a result, the final returned cases can simultaneously cover textual similarity and reflect consistency in issue composition, appeal objectives, and spatiotemporal conditions.

From a methodological perspective, structure-aware fusion and reranking are not a simple addition of two recall results. Instead, they jointly exploit three types of evidence: semantic scores provide cross-expression matching ability, graph scores provide supplementary event-structure information, and cross-encoder reranking re-evaluates the fine-grained alignment between the query and the candidate. Through the collaborative effect of these three components, the system can finally retrieve a high-quality similar-case set that is both semantically relevant and structurally comparable, thereby improving the effectiveness of similar-case retrieval in complex government hotline scenarios.

4. Experimental Setup

4.1. Experimental Datasets and Preprocessing

To evaluate the effectiveness of the proposed dual-route hybrid retrieval method in complex government hotline similar-case retrieval, experiments were conducted on a real government hotline complaint dataset and the public LeCaRD legal case retrieval dataset. The government hotline complaint dataset served as the primary experimental dataset for assessing retrieval performance in a real-world business scenario, while LeCaRD was used as a public benchmark to further examine the cross-domain generalization capability of the proposed method in similar-case matching tasks.

For the government hotline complaint dataset, the original work-order texts were first uniformly cleaned in order to preserve case facts, issue types, and appeal information as much as possible while reducing the influence of noisy text on retrieval experiments. The preprocessing procedure mainly included three steps. First, abnormal characters, repeated spaces, and invalid fields in the original texts were normalized, and privacy-sensitive information was anonymized. Second, a large language model was employed to extract key event elements, including location, time, core issue, and core appeal. Third, the cleaned texts, extracted elements, sample splits, and graph construction results were persistently stored to ensure the reproducibility of the experiments. This preprocessing pipeline provided a unified and reliable data basis for the subsequent similar-case retrieval experiments.

In addition to data cleaning, a relevance annotation scheme tailored to the retrieval objective was constructed. Since similarity between government hotline cases is inherently hierarchical, relevance was not treated as a simple binary label. Instead, a six-level relevance annotation scheme ranging from 0 to 5 was adopted. Specifically, a candidate case was assigned a relevance score of 5 when it was consistent with the query case in first-level category, second-level category, third-level category, handling department, and case title; a score of 4 when the first three category levels and the handling department were consistent; a score of 3 when the three category levels were consistent; a score of 2 when only the first- and second-level categories were consistent; a score of 1 when only the first-level category was consistent; and a score of 0 otherwise. This annotation scheme more finely reflects the hierarchical similarity between government hotline cases and also provides the basis for computing ranking-oriented evaluation metrics such as MAP and Normalized Discounted Cumulative Gain (NDCG).

4.2. Baselines and Evaluation Metrics

To comprehensively assess the effectiveness of the proposed method, multiple representative baseline methods were included in the similar-case retrieval experiments. Traditional lexical retrieval baselines included TextRank, TF-IDF, and BM25, which were used to evaluate the performance of word-matching-based methods in complex hotline work-order scenarios. Semantic retrieval baselines included BERT-based semantic retrieval and DenseRAG based on sentence

embeddings, which were used to compare different semantic recall strategies. In addition, HybridRAG was included as a simple hybrid-recall baseline to examine whether combining multiple retrieval strategies could yield stable gains. The proposed method is denoted as Ours in the experiments.

In terms of evaluation metrics, five commonly used ranking metrics were adopted: P@5, P@10, MAP, NDCG@10, and NDCG@20. Among them, Precision@K (P@K) measures the proportion of relevant cases among the top-K retrieved results and reflects the front-ranking hit capability of the model. MAP evaluates the overall ranking accuracy across different queries. NDCG@K considers both ranking positions and multi-level relevance labels, and is therefore more suitable for the hierarchical similar-case retrieval task considered in this study. Together, these metrics provide a comprehensive evaluation of retrieval performance from both precision-oriented and ranking-quality perspectives.

4.3. Implementation Details

In the proposed method, the retrieval procedure consists of three stages. First, a sentence-embedding model is used to encode both query work orders and historical work orders, and semantic initial recall is performed to obtain semantically relevant candidate cases. Second, graph-based logical recall is conducted on the appeal-driven event knowledge graph to supplement structurally related candidate cases. Third, the candidates returned by the two routes are unified and further reranked by a cross-encoder model, which performs fine-grained interaction modeling between the query and each candidate case to generate the final Top-K retrieval results. Thus, the whole experimental process is organized around the pipeline of semantic recall, graph-based recall, and fusion reranking.

To ensure fair comparison, all baseline methods were evaluated on the same datasets, under the same relevance annotation scheme, and with the same evaluation metrics. Only the retrieval modules themselves were replaced, without introducing additional responsibility-identification or external knowledge-enhancement components. In this way, the influence of non-retrieval factors was minimized, allowing the experimental results to more accurately reflect the relative strengths and weaknesses of different retrieval methods in complex government hotline similar-case retrieval tasks.

5. Experimental Results and Analysis

5.1. Analysis of Similar-Case Retrieval Results

Table 1 reports the similar-case retrieval results on the government hotline complaint dataset. Overall, traditional lexical retrieval methods show relatively limited performance. In particular, TextRank achieves a P@5 of only 0.257, indicating that keyword co-occurrence alone is insufficient for handling the colloquial and fragmented expressions commonly found in government hotline

work orders. BM25 performs better than TF-IDF, which suggests that classical term-statistics-based retrieval models still provide a certain level of utility in this scenario. However, their overall performance remains inadequate for high-quality retrieval in complex hotline cases.

Compared with traditional lexical methods, semantic retrieval methods such as DenseRAG and BERT achieve better results across multiple metrics, indicating that semantic vector representations can effectively alleviate the mismatch caused by different surface expressions. For example, DenseRAG reaches a P@5 of 0.598 and a MAP of 0.620 on the complaint dataset, outperforming lexical baselines and demonstrating the effectiveness of deep semantic features in cross-expression case matching. Nevertheless, such methods still mainly rely on overall textual semantic similarity, and therefore remain limited in modeling internal case structure and event-level consistency. As a result, they may still retrieve cases that are semantically close but not truly comparable in practice.

In contrast, the proposed method achieves the best performance on all evaluation metrics on the complaint dataset. Specifically, it obtains a P@5 of 0.613, a P@10 of 0.582, a MAP of 0.633, an NDCG@10 of 0.659, and an NDCG@20 of 0.722, outperforming all compared methods. These results indicate that, on top of semantic recall, the introduction of graph-based logical recall and structure-aware fusion can substantially improve ranking quality for complex government hotline similar-case retrieval. In particular, compared with DenseRAG, the proposed method improves P@5 and MAP by 0.015 and 0.013, respectively, suggesting that the structural information provided by the event graph effectively compensates for the weakness of pure vector retrieval in modeling case logic.

It is also worth noting that HybridRAG does not show stable performance on the complaint dataset. Its P@5 is 0.494, which is even lower than those of BM25, BERT, and DenseRAG. This suggests that simply combining multiple recall strategies does not necessarily lead to better retrieval performance. Without effective joint modeling and reranking of semantic and structural evidence, hybrid recall may instead introduce additional ranking noise. By contrast, the proposed method does not directly concatenate the outputs of two recall routes, but performs unified ranking through structure-aware fusion and cross-encoder reranking, thereby making the complementary advantages of the dual-route design more effective and stable.

Table 1. Similar-case Retrieval Results on the Complaint Dataset

Method	P@5	P@10	MAP	NDCG@10	NDCG@20
TextRank	0.257	0.222	0.335	0.446	0.555
TF-IDF	0.468	0.440	0.521	0.562	0.647
BM25	0.557	0.528	0.594	0.627	0.695
BERT	0.571	0.535	0.597	0.641	0.708
DenseRAG	0.598	0.568	0.620	0.656	0.719
HybridRAG	0.494	0.522	0.541	0.540	0.633

Our	0.613	0.582	0.633	0.659	0.722

To further evaluate the cross-domain effectiveness of the proposed method, comparative experiments were also conducted on the public LeCaRD legal case retrieval dataset, as shown in Table 2. The results indicate that the proposed method again achieves the best overall performance, with a P@5 of 0.435, a P@10 of 0.395, a MAP of 0.487, an NDCG@10 of 0.754, and an NDCG@20 of 0.807, all of which are higher than those of the baseline methods. This demonstrates that the proposed dual-route hybrid retrieval framework is not only effective in government hotline scenarios, but also exhibits promising generalization ability in more general similar-case retrieval tasks.

Table 2. Similar-case Retrieval Results on the LeCaRD Dataset

Method	P@5	P@10	MAP	NDCG@10	NDCG@20
TextRank	0.308	0.272	0.356	0.608	0.686
TF-IDF	0.340	0.331	0.410	0.153	0.229
BM25	0.377	0.364	0.458	0.720	0.786
BERT	0.088	0.078	0.104	0.145	0.153
DenseRAG	0.349	0.332	0.403	0.680	0.745
HybridRAG	0.379	0.367	0.439	0.721	0.782
Our	0.435	0.395	0.487	0.754	0.807

Overall, the advantages of the proposed method can be summarized from two aspects. First, dense semantic recall preserves the ability to cover cases with substantially different expressions. Second, graph-based logical recall supplements candidate cases with stronger consistency in issue composition, appeal objectives, and event structure. As a result, the final retrieved cases are not only semantically similar, but also structurally comparable, which is particularly important for similar-case retrieval in complex government hotline scenarios.

5.2. Ablation Study and Discussion

To further examine the contribution of each key component in the proposed dual-route hybrid retrieval model, an ablation study was conducted on the complaint dataset. Specifically, three variants were evaluated by removing the semantic matching branch, the graph retrieval branch, and the reranking mechanism, respectively. The results are reported in Table 3. The complete model achieves 0.613, 0.582, 0.633, 0.659, and 0.722 on P@5, P@10, MAP, NDCG@10, and NDCG@20, respectively. When semantic matching is removed, the corresponding values decrease to 0.605, 0.573, 0.629, 0.648, and 0.715. When graph retrieval is removed, they further decrease to 0.598, 0.568, 0.620, 0.656, and 0.719. When the reranking mechanism is removed, performance drops most significantly, with P@5 decreasing to 0.594, MAP to 0.613, and NDCG@20 to 0.708.

Table 3. Ablation Results on the Complaint Dataset

Method	P@5	P@10	MAP	NDCG@10	NDCG@20
Our	0.613	0.582	0.633	0.659	0.722

w/o semantic matching	0.605	0.573	0.629	0.648	0.715
w/o graph retrieval	0.598	0.568	0.620	0.656	0.719
w/o reranking mechanism	0.594	0.552	0.613	0.642	0.708

The ablation results indicate that the cross-encoder reranking mechanism has the largest impact on final retrieval performance. This suggests that although dual-route recall can ensure candidate diversity, it is still difficult to effectively correct ranking errors caused by semantically similar but factually inconsistent cases without fine-grained evidence alignment. In other words, the recall stage mainly determines whether enough useful candidates can be retrieved, while the reranking stage determines whether these candidates can be accurately ordered. Both components are therefore indispensable.

At the same time, removing the graph retrieval branch causes the performance to degrade to a level close to that of DenseRAG, indicating that graph-based logical recall plays an essential role in providing structural evidence beyond semantic matching. Its contribution is not merely adding another recall route; rather, it supplements the retrieval process with information about issue composition, appeal objectives, and spatiotemporal conditions, thereby enhancing structural consistency among candidate cases. Meanwhile, the performance drop caused by removing the semantic matching branch suggests that relying solely on graph structure is also insufficient, especially when event extraction quality is imperfect or graph connectivity is sparse. Therefore, the semantic recall branch remains the basis for ensuring broad and robust retrieval coverage.

In addition to the ablation study, sensitivity analysis was also conducted to examine the variation of P@5 under different relevance-threshold settings. The results show that as the relevance criterion becomes stricter, all compared methods exhibit declining P@5 values. However, the proposed dual-route hybrid retrieval model consistently maintains relatively higher accuracy. In particular, under stricter criteria such as “strongly relevant” and “most relevant,” its performance decreases more slowly than that of lexical retrieval models and single-route semantic retrieval methods. This indicates that, by incorporating graph-structured constraints, the proposed method is more effective in preserving consistency in case facts and complaint appeals, and therefore remains more stable in high-standard similar-case retrieval tasks.

Taken together, the performance gains of the proposed method do not come from any single component, but from the joint effect of semantic recall, graph-based logical recall, and structure-aware reranking. Semantic recall provides cross-expression matching capability, graph-based logical recall enhances structural consistency modeling, and cross-encoder reranking further improves ranking precision. Their collaboration enables the proposed method to achieve more stable and superior performance in similar-case retrieval for complex government hotline scenarios.

6. Conclusion and Future Work

This paper investigates the problem of similar-case retrieval in complex 12345 government hotline scenarios and proposes a dual-route hybrid retrieval method based on an appeal-driven event knowledge graph. To address the limitations of traditional lexical retrieval in handling expression variation and the weakness of purely semantic retrieval in modeling case structure, the proposed method first extracts key elements, including location, time, core issue, and core appeal, from historical work-order texts to construct an appeal-driven event knowledge graph for similar-case retrieval. On this basis, a dual-route hybrid retrieval mechanism integrating dense vector recall and graph-based logical recall is designed, together with structure-aware fusion and cross-encoder reranking for unified candidate ranking. In this way, similar-case retrieval is no longer restricted to superficial textual similarity, but can jointly consider semantic generalization ability and event-structure comparability.

Experimental results show that the proposed method achieves strong performance on both a real government hotline complaint dataset and the public LeCaRD legal case retrieval dataset. On the complaint dataset, the proposed method obtains a P@5 of 0.613 and a MAP of 0.633. On LeCaRD, it achieves a P@5 of 0.435 and a MAP of 0.487, outperforming multiple baseline methods. In addition, the ablation study further demonstrates that semantic recall, graph-based logical recall, and cross-encoder reranking all contribute positively to the final retrieval performance, with the reranking mechanism and the graph retrieval branch playing particularly important roles. These results indicate that the proposed method can effectively improve the accuracy, stability, and structural comparability of similar-case retrieval in complex government hotline scenarios, and can also provide a reliable case basis for higher-level intelligent work-order processing tasks.

Despite these encouraging results, the present study mainly focuses on the similar-case retrieval stage, and does not yet fully explore how retrieved cases can be further used to support practical business decisions. Future work may therefore build upon the proposed high-quality retrieval framework by introducing departmental responsibility knowledge and spatiotemporal contextual knowledge to construct structured decision context for responsible-department identification. In this way, the current retrieval-oriented framework could be extended from identifying comparable historical cases to supporting more intelligent dispatching decisions in complex government hotline scenarios.

In addition, future research may further promote the engineering deployment of the proposed method by developing a prototype system for real-world business processes. Such a system could cover the complete workflow of historical case access, data cleaning, event element extraction, knowledge graph construction, dual-route hybrid retrieval, responsible-department recommendation, and result presentation with manual verification, thereby forming an integrated business loop

for intelligent assistance in government hotline handling.

References

[1] Meng, T.G. and Wang, Y. (2021) Social Governance Innovation Based on Big Data: Local Innovation and Governance Effectiveness. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition), 57, 111-123.

[2] Zhao, S.L. and Wang, J.F. (2020) A Brief Analysis of the Application of Intelligent Government Service Robots. China New Communications, 22, 90.

[3] Zhang, H., Wang, B., Wan, J., et al. (2020) Design and Application of Intelligent Service Robots Based on Artificial Intelligence in the Field of Government Services. Network Security Technology & Application, 3, 122-124.

[4] Aizawa, A. (2003) An Information-Theoretic Perspective of TF-IDF Measures. Information Processing & Management, 39, 45-65.

[5] Robertson, S.E. and Zaragoza, H. (2009) The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3, 333-389.

[6] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.

[7] Reimers, N. and Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982-3992.

[8] Hogan, A., Blomqvist, E., Cochez, M., et al. (2021) Knowledge Graphs. ACM Computing Surveys, 54, 1-37.

[9] Nogueira, R. and Cho, K. (2019) Passage Re-Ranking with BERT. arXiv Preprint arXiv:1901.04085.

[10] Ma, Y., Shao, Y., Wu, Y., et al. (2021) LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2342-2348.

[11] Mihalcea, R. and Tarau, P. (2004) TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 404-411.