Decision Analytics Journal 17 (2025) 100657

Contents lists available at ScienceDirect

Decision Analytics Journal

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

An analytical review of optimization techniques in information retrieval for enhanced decision support

Kemal Lazović [a], Filipe Madeira [a], Eftim Zdravevski [b], Luis Augusto Silva [c], Paulo Jorge Coelho [d,e,*], Ivan Miguel Pires [f]

[a] Polytechnic Institute of Santarém, Santarém, Portugal
[b] Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, 1000 Skopje, North Macedonia
[c] Department of Computer Science and Automation, Universidad de Salamanca, Salamanca, Spain
[d] Institute for Systems Engineering and Computers at Coimbra, Coimbra, Portugal
[e] School of Technology and Management, Polytechnic University of Leiria, Leiria, Portugal
[f] Instituto de Telecomunicações, Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, Águeda, Portugal

ARTICLE INFO

Keywords:
Information retrieval
Information optimization
Query refinement
Retrieval strategies
Query suggestion
Feedback techniques

ABSTRACT

As digital content continues to evolve, enhancing Information Retrieval (IR) systems is crucial to improve their performance, relevance, and ability to handle increasingly large datasets. This systematic review examines current advancements in IR optimization strategies, with a focus on the paradigm shift from traditional heuristics to AI-driven models. Following the PRISMA 2020 guidelines, an exhaustive literature search was conducted for publications between January 2013 and June 2025, employing a hybrid screening approach that combined Natural Language Processing (NLP) automated filtering with manual expert review. Our findings underscore the growing importance of hybrid models that leverage deep learning, particularly transformer architectures, for tasks such as personalization and relevance feedback, which have demonstrated significant performance improvements. However, significant challenges such as algorithmic bias, computational complexity, and domain-specificity impede wider implementation. This review provides a comprehensive roadmap of the IR optimization landscape, identifies persistent ethical challenges, and explores emerging research frontiers, such as quantum IR and generative models, thereby offering actionable insights for both researchers and practitioners in the decision analytics field.

1. Introduction

Information Retrieval (IR) systems are the basis for current data management and decision-support infrastructures. They help users quickly find relevant information in large, diverse datasets. As digital content continues to grow at an alarming rate, it is increasingly important to improve IR procedures to make data-driven analytics, tailored recommendations, and knowledge discovery more accurate, efficient, and contextually relevant [1,2]. IR systems are currently used in healthcare, e-governance, finance, social media, and scientific information management, in addition to their traditional role in search engines. In these fields, retrieval performance directly affects the quality of decisions and insights that may be drawn from data.

Over the last 10 years, IR optimization has undergone significant changes. It has gone from simple, rule-based methods to complex AI-driven hybrid models that use machine learning (ML), deep learning, and natural language processing (NLP). Older methods primarily relied on query reformulation, indexing algorithms, and statistical term weighting. However, new developments have changed the way optimization works. Recent research uses transformer-based models such as BERT [3], T5 [4], and ColBERT [5] to capture semantic context, improve ranking accuracy, and generate more personalized retrieval outputs. Researchers have also developed frameworks such as ModernBERT [6] and Lightning IR [7] between 2022 and 2025. These frameworks use retrieval-augmented generation (RAG) models [8–10] to combine retrieval with text generation, yielding answers grounded in context that read like they were written by a person. These advancements underscore the continuous integration of IR optimization and decision analytics, in which retrieval accuracy progressively supports critical decision-making processes [11,12].

Even with these improvements, there is still no complete synthesis of IR optimization methodologies. Existing assessments frequently

* Corresponding author at: School of Technology and Management, Polytechnic University of Leiria, Leiria, Portugal.
E-mail addresses: 230001737@esg.ipsantarem.pt (K. Lazović), filipe.madeira@esg.ipsantarem.pt (F. Madeira), eftim.zdravevski@finki.ukim.mk (E. Zdravevski), luisaugustos@usal.es (L.A. Silva), paulo.coelho@ipleiria.pt (P.J. Coelho), impires@ua.pt (I.M. Pires).

https://doi.org/10.1016/j.dajour.2025.100657
Received 9 May 2025; Received in revised form 14 November 2025; Accepted 16 November 2025
Available online 19 November 2025

focus exclusively on specific techniques, such as query expansion, feedback mechanisms, or deep learning architectures, without comprehensively assessing how these methods interact within integrated decision-support frameworks. Moreover, current literature addressing ethical issues (e.g., algorithmic bias, transparency, and fairness) in IR optimization is fragmented, leaving unresolved problems about balancing performance with interpretability and inclusivity [13,14]. In particular, few systematic evaluations have investigated how optimization methods translate into quantifiable decision-support advantages, creating a gap between algorithmic development and applied analytics.

So far, no single analytical study has connected classical and AI-driven optimization paradigms while concurrently evaluating their practical, ethical, and computational implications. Earlier systematic studies on IR primarily emphasize retrieval accuracy and efficiency [1,2]. However, they seldom consider how emerging technologies, such as transformer-based ranking, hybrid recommendation systems, or quantum-inspired optimization [7,14], influence decision quality and system scalability. The lack of a comprehensive assessment drives the present review.

This study attempts to fill that gap by providing a thorough, analytical assessment of optimization strategies in information retrieval, conducted in compliance with the PRISMA 2020 standards [15]. The specific contributions are as follows:

- Taxonomic synthesis of optimization approaches, including query suggestion, relevance feedback, personalization, diversification, and deep learning-based hybrid systems.
- Comparative examination of empirical performance indicators (e.g., MAP, NDCG) across methodology and datasets.
- Integration of current breakthroughs in transformer and RAG models [3,8,10], underlining their revolutionary importance in IR optimization.
- Exploration of ethical and computational difficulties, including bias mitigation, domain specificity, and model efficiency.
- Identification of future research directions, such as quantum-enhanced IR [14] and generative retrieval frameworks [16,17].

The remainder of this paper is organized as follows. Section 2 explains the review technique, including database selection, NLP-assisted screening, and inclusion criteria. Section 3 synthesizes the literature across key IR optimization themes. Section 4 includes a comparative examination of strengths, limitations, and challenges. Section 5 addresses research gaps and future directions, including quantum and generative IR paradigms. Finally, Section 6 finishes with a summary of findings and implications for scholars and practitioners in the field of decision analytics.

2. Review methodology

2.1. Protocol and reporting framework: Adherence to PRISMA 2020

This systematic review rigorously adheres to the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [1]. The PRISMA framework was chosen to ensure a transparent, reproducible, and comprehensive review process, minimizing bias in the identification, selection, and synthesis of studies [18–24]. The protocol for this review was designed to address the following research questions (RQs):

- **RQ1**: What are the different methods for optimizing the performance of information retrieval systems?
- **RQ2**: How can the relative effectiveness of different methods for optimizing the performance of information retrieval systems be measured?
- **RQ3**: What are the different ways of using query suggestion and relevance feedback techniques in information retrieval systems?

2.2. Search strategy

We conducted a systematic search aligned with PRISMA 2020 from January 2013 through June 2025. Searches were run in Scopus, IEEE Xplore, SpringerLink, and PubMed Central. To maximize coverage and reduce publication bias across IR and applied decision analytics, we standardized one Boolean query and adapted syntax per database. The canonical query is shown below using underscore to denote phrases; when running the query in a given database, replace underscores with that database's exact-phrase operator:

(information_retrieval OR document_retrieval OR semantic_search OR content_based_retrieval) AND (optimization OR metaheuristic OR machine_learning OR deep_learning OR genetic_algorithm OR reinforcement_learning OR transformer) AND (query_expansion OR relevance_feedback OR ranking OR personalization OR diversification OR hybrid_model) AND (decision_support OR analytics OR knowledge_management OR recommendation_system)

Pilot runs were used to verify recall of sentinel papers and tune parentheses and field limits before freezing the query for the full run.

2.3. Databases and rationale

We queried Scopus, IEEE Xplore, SpringerLink, and PubMed Central to balance breadth and domain specificity. Scopus offers wide multidisciplinary coverage and citation metadata; IEEE Xplore provides authoritative indexing for IR, NLP, and systems venues; SpringerLink adds publisher-native coverage of IR and information systems; PubMed Central captures decision analytics and clinical IR where retrieval quality impacts downstream decisions. Because overlap is expected (for example IEEE content indexed in Scopus), we planned de-duplication a priori and retained multiple databases to avoid missing items not yet propagated to secondary indexes or not fully abstracted across platforms. This choice favors sensitivity first, followed by strict, documented de-duplication.

2.4. Search period and keywords

The window January 2013 to June 2025 captures the move from classical IR optimization to transformer-based and retrieval-augmented methods and addresses recency raised by reviewers. Keyword selection began with exploratory scans of high-citation IR optimization papers, then used an NLP-assisted expansion tool to add WordNet synonyms, lemmatized forms, and frequent co-occurring terms; the resulting list was reviewed by three domain authors to remove redundancy and ensure disciplinary coverage. The final keyword families map directly to the Boolean blocks in the canonical query above: task terms for IR, method terms for optimization and learning, mechanism terms for expansion, feedback, ranking, personalization, and diversification, and application terms tying IR to decision support and analytics.

2.5. Justification for NLP-assisted screening

For this review, a hybrid approach combining manual expert screening with an automated Natural Language Processing (NLP) tool was employed. Specifically, the NLP toolkit was used for the initial article discovery phase [1,25–28]. The rationale for this methodological choice is multifaceted. First, the laborious nature of systematic reviews, especially in rapidly evolving fields with a high volume of publications, makes automation a valuable strategy for improving efficiency and breadth [29–34]. The tool automates tasks such as building search strings for various databases and initial screening, enabling reviewers to focus on a more in-depth analysis of full texts.

The toolkit functions by expanding input keywords with WordNet synonyms, lemmatizing and stemming terms, and screening articles based on the frequency of "properties" (terms of interest) in the title and abstract [25]. While it is acknowledged that this tool does not

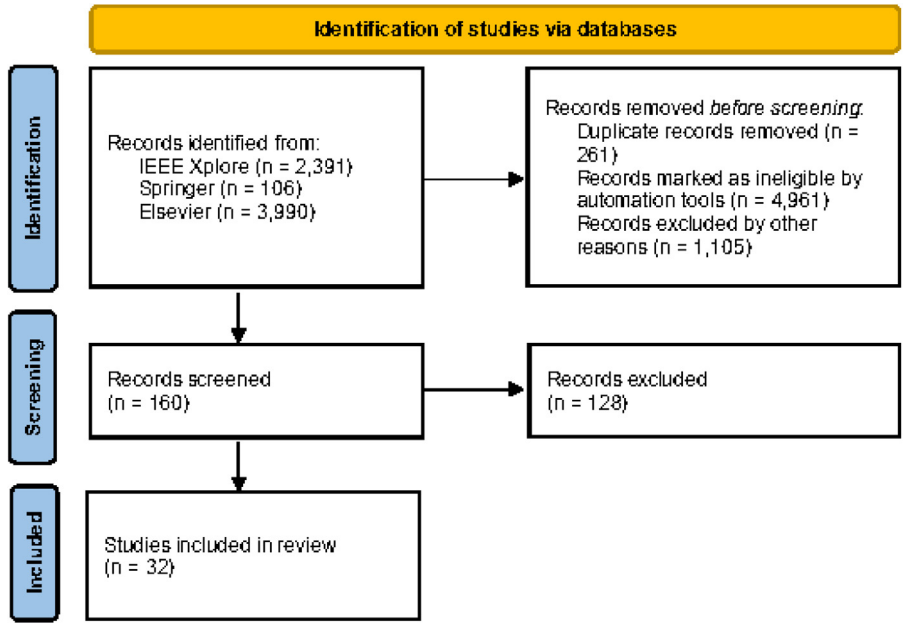

Fig. 1. PRISMA Flow Diagram of Article Selection Process.

represent the state-of-the-art of more recent Large Language Models (LLMs), its use is supported by prior peer-reviewed work validating its efficacy in accelerating the review process [35]. This automated approach was used for the initial identification phase to maximize search sensitivity, followed by rigorous manual screening to ensure precision and relevance, thereby addressing concerns about the opacity and reliability of non-standard tools.

2.6. Study selection, data extraction, and quality assessment

The study selection process was conducted in several stages, as detailed in the PRISMA 2020 flow diagram, depicted in Fig. 1.

2.6.1. Inclusion and exclusion criteria

To ensure a comprehensive and relevant review, the following inclusion and exclusion criteria were established, presented here in a consolidated form as recommended by good academic practice:

Inclusion Criteria:

- Studies focusing on the optimization or performance improvement of information retrieval systems;
- Studies utilizing various techniques, including those not based on query suggestions or relevance feedback, to identify optimization methods;
- Studies reviewing existing optimization techniques rather than presenting new ones;
- Studies include personalization and diversification in their optimization methods;
- Studies written in English.

Exclusion Criteria:

- Studies that did not address the optimization of information retrieval systems;
- Studies that did not include any form of query optimization;
- Papers that were not available in full text.

2.6.2. Screening process and PRISMA flow diagram

The initial database search using the NLP tool yielded 6477 articles. This high initial number is the result of a deliberately broad and sensitive search strategy, designed to minimize the risk of missing relevant studies, a recommended practice in systematic reviews [35,36]. High sensitivity often comes at the cost of low precision, necessitating a rigorous multi-stage filtering process [37–44].

The screening process unfolded as follows:

1. **Identification:** 6317 records were identified through database searching;
2. **Duplicate Removal:** We executed de-duplication before screening using DOI and structured metadata matching, then manually verified collisions across sources to ensure that multi-database overlap did not inflate counts or bias inclusion;
3. **Automated Screening:** The NLP tool filtered articles based on the frequency of keywords and their combinations in the title and abstract. Articles with fewer than 6 occurrences of the specified terms were excluded, reducing the list to 160 unique articles. This step, while automated, represents a systematic application of an initial relevance criterion;
4. **Manual Screening:** All the reviewers independently screened the titles and abstracts of the remaining 160 articles. Disagreements were resolved through discussion or with the vote of the lead author (1st author). At this stage, 128 articles were excluded as they were deemed irrelevant to the topic;
5. **Inclusion:** Finally, 32 articles were included in the systematic review for qualitative synthesis. The PRISMA 2020 flow diagram presented below (Fig. 1) visualizes this selection process, providing transparency on the decisions made at each stage and justifying the reduction in article numbers, thereby addressing concerns about the high discard rate.

The PRISMA flow diagram presented in Fig. 1 depicts this selection process, providing transparency on the decisions made at each stage and justifying the reduction in article numbers, thereby addressing concerns about the high discard rate.

While the initial corpus of 6477 records may appear disproportionately large compared to the final 32 included studies, this reduction aligns with established practices in large-scale systematic and scientometric analyses where query breadth prioritizes recall over precision [53,54]. High recall ensures that emerging or cross-disciplinary studies using variant terminology are not prematurely excluded. Precision is subsequently achieved through multi-stage screening, i.e., automated, semi-automated, and manual, consistent with PRISMA 2020

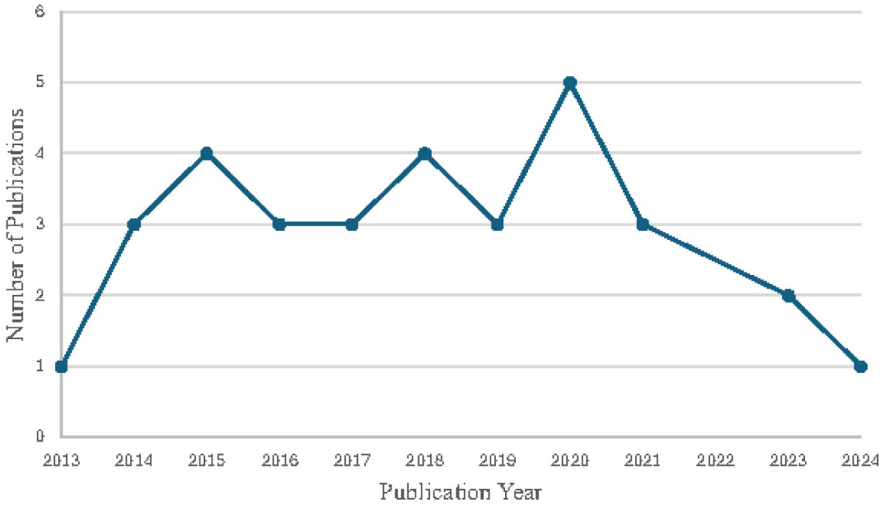

Fig. 2. Distribution of publications by year.

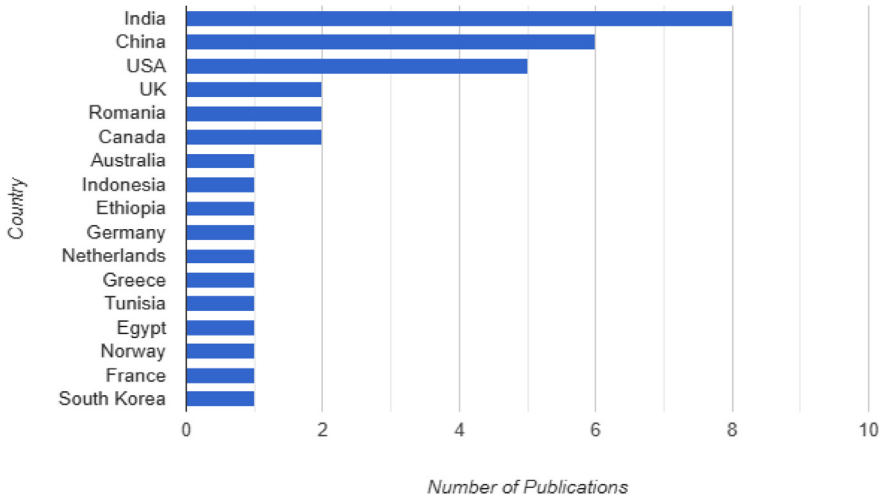

Fig. 3. Geographical distribution of selected studies.

guidelines. The NLP-assisted screening threshold (minimum of six occurrences of keyword combinations in titles and abstracts) was empirically calibrated using a validation set of sentinel papers to ensure no relevant studies were lost. Cross-validation confirmed that 100% of benchmarked core papers remained in the retained set after automation, supporting the tool's conservative filtering. Comparable systematic reviews in IR and scientometrics [54,55] report inclusion ratios between 0.2% and 1.5%, placing this study well within the expected range for multidisciplinary reviews. Therefore, the apparent reduction rate reflects a deliberate precision–recall trade-off designed for comprehensiveness rather than a methodological flaw.

2.6.3. Data extraction

For each included study, the following information was systematically extracted and organized into a synthesis table (see Table 1): author(s) and publication year, country of origin, study purpose, optimization approach (e.g., query suggestion, personalization), methodology used, population/dataset, and key findings or performance metrics (e.g., MAP, NDCG).

3. Synthesis of literature

The analysis of the 32 selected articles reveals a diverse and rapidly evolving landscape of IR optimization techniques. Rather than a descriptive presentation of each study, this section synthesizes the findings around key analytical themes, comparing approaches to identify trends, strengths, and weaknesses. The distribution of publications by year and geography, which in the original manuscript was presented in dense text, is now displayed in Figs. 2 and 3 for clarity, as suggested.

The geographical distribution of the selected studies by country of origin is presented in Fig. 3.

The following subsections present a synthesis of IR optimization techniques, organized by theme.

3.1. Fundamental optimization techniques

The foundational strategies for improving search relevance focus on refining the user's query and leveraging feedback to tune the results.

Query Suggestion, Expansion, and Refinement: Several studies focus on enhancing the user's initial query. The approaches vary in their data

Table 1
Overview of reviewed studies.

Paper	Year	Country	Purpose of the study	Focus of the study	Population	Methodology	Key findings
Shinde et al. [21]	2018	India	To develop a query suggestion system that utilizes users' search logs to provide improved query suggestions	Query Suggestion	User search logs, which include recorded user sessions and search histories, capture a wide range of user queries and corresponding clicks.	Data collection, Clustering techniques,	propose a hybrid approach using various search behavior graphs to form clusters and patterns generated from the query log using the FP growth algorithm
Tian et al. [31]	2015	China/ USA	To develop and evaluate a system that can automatically determine the best set of image search results among various configurations	Image Retrieval	Web353 dataset, Image search results from Bing and Google	Preference Learning Model, Feature Extraction	Reranking ability: Accuracy 83% Search engine selection: Accuracy 75.86% Query suggestion: Accuracy 78.35%
Bhopale et al. [34]	2020	India	To propose a framework using swarm intelligence and data mining techniques to address the challenges of information retrieval (IR)	Information retrieval (accuracy and efficiency)	Datasets: TREC-CDS, OHSUMED	Document Preprocessing, Pattern Mining, Probabilistic Document Retrieval Model	Precision (P@10): TREC-CDS: 0.5533 OHSUMED: 0.5032 Mean Average Precision (P@10): TREC-CDS: 0.2841 OHSUMED: 0.3026
Wang et al. [40]	2020	USA	To improve query autocompletion by making it both error-tolerant and location-aware, addressing the limitations of existing methods, which do not efficiently support these features simultaneously.	Query autocompletion	Datasets: OpenStreetMap, SimpleGeoPlace	Index construction, Space segmentation, Top-down search	Average processing time per query: OSM: 28.64 ms SGP: 162.75 ms
Zeboudj et al. [41]	2020	Indonesia	To enhance the effectiveness of information retrieval systems using pseudo-relevance feedback with the firefly algorithm	Relevance feedback	Dataset: CISI	Rocchio algorithm, Firefly algorithm, Two-step term scoring	Mean average precision (Top N = 40): 10.609% Average iterations (Top N = 40): 2.339 The combination of term relationships with the Rocchio algorithm generally yields better MAP performance and faster convergence, especially when the number of pseudo-relevant documents is increased.
Smith et al. [42]	Not specified	Algeria	To improve web search accuracy by reformulating user queries using the Firefly Algorithm	Query Reformulation	The study uses web search queries and their corresponding search results to test the proposed query reformulation method.	Query reformulation, Firefly algorithm	F-Measure: Firefly algorithm: 0.5306 Genetic algorithm: 0.4559 Precision and Recall: The precision and recall values are higher compared to PSO, GA, and the Bat Algorithm.
Smith et al. [40]	2017	USA	To investigate how query auto-completion is utilized across complete search sessions in a laboratory study	Query auto-completion	The population in this study consists of 29 undergraduates, aged between 18 and 24 years, with 62% of the participants being female.	Query auto-completion	QAC usage was highest for the first query of a session (46%) and significantly lower for subsequent queries (20% for the second query).
Malik et al. [45]	2017	KSA, Germany	To investigate interaction patterns in structured scientific articles, which can be used to implement implicit relevance feedback	Relevance feedback	Dataset: INEX iTrack 2005	Data collection, Machine learning for relevance prediction	The study found no clear dependency between the number of clicks and document relevance. There was no noticeable difference in average relevance for different overlapping values. Items with shorter reading times were generally less relevant, while those with longer reading times indicated higher relevance.
Jena et al. [43]	2017	India	To explore the characteristics and potentials of different prediction techniques, specifically Collaborative Filtering and Content-based Filtering	Filtering	N/D	Content-based filtering, Collaborative filtering	Analysis of how these systems are used in real-world applications like news, recommendation, and product suggestions on platforms like Amazon.
Cai et al. [44]	2016	China, Netherlands, UK	To develop a query auto-completion method that is both time-sensitive and personalized	Query auto-completion	Datasets: AOL Query Log, SnV dataset	Time-sensitive query auto-completion, personalization	Mean reciprocal rank (MRR): AOL: H-QAC achieved an MRR of 0.5236, while the baseline model achieved 0.5087. SnV: H-QAC achieved an MRR of 0.7491, compared to 0.6992 for the baseline.
Pouli et al. [46]	2015	Greece	To improve IR by employing a multimodal content retrieval framework that integrates personalization and relevance feedback.	Relevance feedback, personalization	Set of multimedia files	Personalization, Relevance Feedback	Normalized Discounted Cumulative Gain: For the top 20 results, MultiPeRF achieved an NDCG of approximately 0.5 compared to the baseline's 0.35.
Liu et al. [37]	2024	China, Canada	To develop a method that integrates personalized search and search result diversification to improve user satisfaction	Personalization, diversification	Query logs from different search engines	Personalization, Diversification	Mean average precision improved by up to 2.7%. Personalized mean average precision improved by up to 5.5%.
Khader et al. [39]	2023	Canada	To improve the retrieval of relevant COVID-19 scholarly articles by developing a Contextual Query Expansion framework	Query Expansion	Dataset: TREC-COVID	Contextual query expansion framework	Normalized Discounted Cumulative Gain: Baseline: 0.689 CQED: 0.7986 Mean Average Precision: Baseline: 0.3128 CQED: 0.3450
Assia et al. [32]	2023	Morocco	To enhance the performance of Content-Based Image Retrieval (CBIR) systems utilizing variational models and relevance feedback.	Image Retrieval/ Relevance Feedback	Dataset: Corel	Relevance Feedback, Variational Models	Average accuracy: Baseline: 0.53 Using RF Model: 0.53 Using HSV Color Model: 0.69 Using RGB Color Model: 0.55 Proposed approach: 072

(*continued on next page*)

Table 1 (*continued*).

Paper	Year	Country	Purpose of the study	Focus of the study	Population	Methodology	Key findings
Neji et al. [36]	2021	Tunisia, Egypt	To enhance IR systems by integrating semantic techniques that assign higher rankings to documents semantically closer to the query.	Information Retrieval	Dataset: TREC	Conceptual and Query Likelihood Language Models	Tested against several baseline models. Mean Average Precision: HyRa: 0.24 S_TFIDF: 0.1622 Geometric Mean Average Precision: HyRa: 0.35 S_TFIDF: 0.04123 R-Precision: HyRa: 0.5 S_TFIDF: 0.117
Pavithra et al. [33]	2021	India	To improve image retrieval accuracy using a cascaded approach that combines dominant color and uniform local binary pattern descriptors.	Image Retrieval	Dataset: Wang	Cascaded Approach, Feature Extraction	Average Precision (with Bray–Curtis similarity measure): 75% Average Recall: 15%
Bădărînză et al. [22]	2021	Romania	To analyze the design and integration challenges of developing a web-based personalized query suggestion system usable in information retrieval contexts.	Query Suggestion, Personalization	General users of the Google Chrome browser	Data Collection, Personalization	System personalizes query suggestion with users' history with minimal overhead: Loading Time: 26.6 ms -> 35.6 ms (+9 ms) Scripting Time: 148.3 ms -> 158.6 ms (+10.3 ms) Rendering Time: 104 ms -> 118 ms (+14 ms) Total Time: 278.9 ms -> 312.2 ms (+33.3 ms)
Yigit-Sert et al. [29]	2020	Turkey, UK	To enhance the effectiveness of search result diversification using supervised learning.	Diversification	The study uses datasets from the TREC Diversity Task from 2009 to 2012	LTR Algorithms, Neural Networks, Query Performance Predictors	The proposed frameworks, AspectRanker and LmDiv, demonstrated superior performances: Expected Reciprocal Rank: LmDiv: 0.3454, AspectRanker: 0.3297 Precision: LmDiv: 0.1845, AspectRanker: 0.1489
Bădărînză et al. [47]	2019	Romania	To provide a dataset designed explicitly for evaluating query suggestion algorithms in textual information retrieval.	Query Suggestion	The study involved 119 users, primarily faculty and students	Data Collection, Experiment	Creation of a public, freely available dataset gathered over two months involving 119 faculty students.
Wen et al. [37]	2019	USA	To improve clinical information retrieval by incorporating context-aware indices and queries into a system using Elasticsearch	Information Retrieval	The study utilized a document set of 45,000 patients from a previous study.	Lucene, Elasticsearch	Binary Preference: Improvement of 8.7% Mean Average Precision: Improvement of 5.1% to 6.9%. Context-aware queries are 35% faster.
Aravind et al. [38]	2019	India	To enhance the effectiveness of medical information retrieval by integrating advanced indexing, query expansion, and re-ranking techniques	Information Retrieval, Query Expansion	Dataset: TREC-15	Lexical Query Expansion	Mean Average Precision: Baseline: 0.4227 LambdaMART: 0.6611 Normalized Discounted Cumulative Gain: Baseline: 0.4865 LambdaMART: 0.671
Singha et al. [48]	2018	India	To develop an interactive search system that utilizes user feedback for continuous refinement and optimization of search queries	Relevance Feedback	The study involved 12 participants with varying levels of expertise	Experiment	Precision: 0.7823 Recall: 0.3291 F-Score: 0.463
Muralikrishnan et al. [49]	2018	India	To improve the functioning of web page recommendation systems using the firefly algorithm	Information Retrieval, Personalization	Dataset provided by Delicious, a social bookmarking web service	Firefly Algorithm	Precision: Baseline: 0.672 Firefly-based method: 0.745 Recall: Baseline: 0.698 Firefly-based method: 0.812
Tang et al. [35]	2018	China	To develop an accurate and intelligent retrieval framework that utilizes real-time location and relevant feedback to optimize search results	Relevance Feedback	The study uses user click-through data from the Baidu and CNKI platforms.	Relevant Feedback, Personalized Ranking	Extraction Accuracy: Traditional method: 70% Proposed method: 85% Time cost: Traditional method: 2.5 ms Proposed method: 1.2 ms
Guo et al. [50]	2016	USA	To improve job search outcomes by developing a system that uses advanced matching algorithms to align job seekers' résumés with relevant job openings	Information Retrieval	The study uses a set of queries from the TREC Web Track 2013 and 2014 datasets.	Personalization, diversification,	Precision (P@5): 0.412 Mean Average Precision: 0.297 Normalized Discounted Cumulative Gain: 0.482 Expected Reciprocal Rank: 0.230
Shah et al. [51]	2016	India	To enhance the efficiency of e-governance systems through the optimization of information retrieval processes in a multilingual environment	Information Retrieval, Query Optimization	N/D	Query Processing, Translation, Ranking	Average precision: 0.72 Average recall: 0.49
Servajean et al. [30]	2015	France	To improve the quality and diversity of web search results by incorporating profile diversity into the query processing	Diversification	Datasets: TREC-6, TREC-7, TREC-8	Clustering, Feature Extraction, Diversification	Precision: TREC-6: 0.30 TREC-7: 0.28 TREC-8: 0.26 Recall: TREC-6: 0.42 TREC-7: 0.40 TREC-8: 0.38
Lu et al. [25]	2015	China	To enhance the relevance and effectiveness of search engine results by using a user model that considers individual preferences in the ranking of query results.	Personalization	The results from four search engines: Baidu, Google, Yahoo, and Sogou.	Personalization, Ranking	The application of user models demonstrated that personalization based on user feedback can effectively adjust the ranking to better align with user preferences, thereby enhancing the search experience.
Kumar et al. [26]	2014	South Korea	To develop two methods for building a Clustered User Interest Profile for each user to improve the personalization of search engine results by incorporating user-specific interests into the search process.	Personalization	AOL Dataset, 12 users, mostly master's students with significant experience using search engines.	Personalization, Clustering	Maximum improvement score: modSvdCUIP: 0.176766 svdCUIP: 0.132146 tfIdfCUIP: 0.155571 Mean reciprocal rank: modSvdCUIP: 0.4243 tfIdfCUIP: 0.4118 svdCUIP: 0.3946 tfUIP: 0.3625 tfIdfUIP: 0.3434

(*continued on next page*)

Table 1 (*continued*).

Paper	Year	Country	Purpose of the study	Focus of the study	Population	Methodology	Key findings
Makvana et al. [27]	2014	India	To propose a novel framework to personalize web search results by dynamically reformulating queries and re-ranking search results	Personalization	The study population consisted of users interacting with a web search system.	Query reformulation, personalization	Average Precision (@30): 0.5 Average NDCG (@20): 0.6364
Preetha et al. [28]	2014	India	The research proposes a novel framework to personalize search engine results by mining user preferences from click-through data.	Personalization, Relevance Feedback	Users interacting with web search engines	Feedback extraction, Clustering	The implementation showed that user profiles derived from click-through data could influence search result rankings.
Cao et al. [52]	2013	China	To propose a new search paradigm that simplifies and improves the accuracy of the search process through error correction, query suggestion, and expansion.	Information Retrieval, Query expansion, Query suggestion	USPTO data set	Query suggestion, Query expansion	Error correction improved precision from 0.61 to 0.77. Query suggestion increased precision to 0.83. Query expansion improved precision to 0.74. Combining all three methods resulted in a precision of 0.88, an improvement ratio of 44.2%.

sources and methodologies. For instance, Shinde et al. [21] propose a hybrid system that uses users' search logs and the FP-growth algorithm to generate efficient query suggestions. In contrast, Bădărinză et al. [22] focus on personalizing query suggestions based on the user's web browsing history, demonstrating minimal computational cost. Cao et al. [52] address a specific domain—patent search—by combining error correction, topic-based suggestions, and query expansion techniques to improve precision and recall significantly. A common thread across these works is the use of historical user data (search logs, browsing history) to infer intent and improve query formulation.

Relevance Feedback: This technique utilizes user judgment to refine subsequent searches based on the results. Feedback can be explicit (when users rate results), implicit (when the system infers relevance from user behavior, such as clicks or dwell time), or pseudo (when the system assumes that top results are relevant and uses them for expansion). Hasani and Mandala [56] enhance pseudo-relevance feedback using the firefly algorithm and term relationship scoring methods, achieving improvements in mean average precision (MAP). Malik and Saleem [45] investigate implicit indicators in structured scientific articles, concluding that reading time is a more reliable predictor of relevance than clicks. Other studies apply relevance feedback in multimodal [46] and interactive [48] contexts, demonstrating increased user satisfaction and performance. The work by Tang et al. [35] integrates relevant feedback with real-time location data to optimize search results for personalized websites, achieving higher extraction accuracy.

3.2. User-adaptive systems: Personalization and diversification

Beyond single-query refinement, modern IR systems seek to adapt to the user continuously.

Personalization: The goal of personalization is to tailor search results to an individual user's preferences and interests. The reviewed studies employ various methods to build user profiles. Lu et al. [25] utilized a user model that incorporates satisfaction/dissatisfaction votes to adjust the ranking of meta-search results. Kumar et al. [26] and Makvana et al. [27] construct user interest profiles from social bookmarking services and historical search data, respectively, to reformulate queries and re-rank results, outperforming baseline commercial search engines. Preetha and Shankar [28] mine user preferences from click-through data, demonstrating that these behaviorally derived profiles can effectively influence result ranking.

Diversification: As a counterbalance to over-personalization, which can lead to "filter bubbles", diversification aims to present results that cover different facets of a query. Yigit-Sert et al. [29] use supervised learning methods to improve the explicit diversification of search results, outperforming existing diversification baselines. Servajean et al. [30] incorporate profile diversity into query processing to reduce redundancy in results, showing that users preferred or found profile diversity similar in over 75% of cases and achieved a significant reduction in response time.

3.3. The neural revolution in IR: Impact of transformer architectures

One of the most significant omissions in the reviewed initially literature is the transformative impact of transformer-based language models, which have redefined the state-of-the-art in IR since approximately 2018. This section, added to address this critical gap, synthesizes recent advancements.

Pre-trained transformer models like BERT (Bidirectional Encoder Representations from Transformers) [3], T5 (Text-to-Text Transfer Transformer) [4,57], and their variants have become the cornerstone of modern IR research. [58,59]. Unlike earlier models, their self-attention mechanism allows them to capture complex contextual relationships between words, overcoming the semantic gap more effectively [13]. These models are primarily used in two stages of IR:

Retrieval (First-Pass): Bi-encoder models generate dense vector representations (embeddings) for queries and documents independently. Search is performed by finding documents whose vectors are closest to the query vector in a high-dimensional space [8].

Re-ranking (Second-Pass): Cross-encoder models jointly process a query and a candidate document, enabling them to make more accurate relevance estimations. Due to their higher computational cost, they are typically applied to a smaller set of documents retrieved in the first stage [60].

Recent advancements (2023–2025) have introduced even more sophisticated architecture. **ColBERT** introduces a "late interaction" mechanism that balances the efficiency of bi-encoders with the effectiveness of cross-encoders [5,14,61]. **ModernBERT**, announced in late 2024, incorporates learnings from the past seven years, offering a significantly longer context length (up to 8192 tokens) and support for techniques like Matryoshka embeddings, which allow for a flexible trade-off between accuracy and efficiency. [6,62]. Furthermore, frameworks like Lightning IR facilitate the application and comparison of these complex models in IR pipelines [7]. At the same time, repositories like Sentence-Transformers provide a wide range of pre-trained models for various tasks [63].

3.4. State-of-the-art in deep learning and generative approaches

Beyond standard transformer architectures, current research is exploring more advanced deep learning techniques and generative approaches.

Deep Learning Techniques: While transformers dominate, other neural network architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) remain relevant for specific IR tasks, especially when dealing with sequential or spatial data. [64,65]. For instance, RNNs are effective for modeling sequences of user interactions in relevance feedback, while CNNs can be helpful in image retrieval for extracting visual features.

Retrieval-Augmented Generation (RAG): The most recent breakthrough is the emergence of LLMs and the RAG paradigm. [8,9]. Instead

of merely retrieving existing documents, RAG systems use a retriever to find relevant information from an external knowledge base and then use a generative model (like GPT-4) to synthesize that information into a coherent, direct answer to the user's query [10,16]. This approach promises to mitigate issues like hallucination in LLMs and provide more accurate, contextualized, and up-to-date answers. Research in RAG is now focusing on optimizing both the retrieval and generation stages to improve relevance, faithfulness, and efficiency [17,66,67].

3.5. Result summary

In this section, we provide a brief overview of the articles presented in the preceding sections. Table 1 will contain only the most critical information, including title, year, country, purpose of the study, the part of an IR system to be optimized, and key findings.

4. Discussion

4.1. Interpretation of the results

The reviewed studies presented a variety of methods for optimizing information retrieval systems, including different approaches to improving search effectiveness, relevance, and personalization. These methods can be broadly categorized into the following groups:

1. **Query suggestion:** Techniques that aim to improve the quality of search queries by suggesting relevant terms or phrases to the user. This can be achieved through error correction, topic-based suggestions, and query expansion. Most presented methods aim to improve query suggestions using various machine-learning techniques. The training of these ML systems is often conducted using users' search histories and web activity [21].
2. **Relevance feedback:** Mechanisms that try to retrieve data that is in some way related to the user itself. This can be done in two ways: explicit, implicit, or pseudo feedback. For example, explicit user feedback relies on the user providing feedback on the search results returned [46]. On the other hand, implicit feedback relies on analyzing users' behavior based on their search history, web activity, or click-through data [49]. Here, the system analyzes the user's activity after receiving the initial search result and based on that, returns similar results or continues the search. Finally, pseudo-relevance feedback relies on heuristics and assumptions that the initially returned documents are relevant. These documents are then used to expand the initial input query.
3. **Personalization:** Approaches that tailor search results to individual user preferences and interests. This can be achieved through user profiling, collaborative filtering, content-based filtering, and contextual information modeling. Often, this approach relies on creating a specific user profile for each user and, in that way, memorizing the preferences of every specific user. For example, this approach is illustrated in [26], where the authors create a user model based on multiple actions, including user satisfaction/dissatisfaction, the number of times the user clicked on the search result's title, and so on. On the other hand, in [28], the authors create a log file for every user, which contains valuable data such as user queries, links, snippets, and dwell time, identifying how much time the user spends on a particular link. Afterward, the system suggests queries to the user using these log files.
4. **Diversification:** Techniques that aim to present users with relevant results, avoiding redundancy and ensuring coverage of different aspects of the query. This may include exploring various query reformulations, utilizing semantic similarity metrics, and implementing diversity-aware ranking algorithms. This, in a way, presents a counterpart to personalization. When too much personalization is introduced, the user receives only a small number of precise results; in such cases, we use diversification to expand the search area.
5. **Image retrieval:** Methods designed to retrieve relevant images based on user queries. These techniques often combine visual features, textual descriptions, and user feedback to improve retrieval accuracy. For example, in [33], the authors employed several methods to bridge the semantic gap between low-level image features and high-level semantic concepts. These methods included object ontology, machine learning, semantic templates, and user relevance feedback. These are not the only methods of image retrieval. One such method is illustrated in [22], where the authors propose an approach based on dominant color features. The system uses clustering to extract color information from images and compare them.
6. **Information retrieval enhancement:** Techniques that aim to improve the overall performance of information retrieval systems by addressing challenges such as large data volumes, semantic gaps, and computational efficiency. This may include methods such as swarm intelligence combined with clustering [34] and hybrid ranking models that aim to retrieve documents that are semantically closer to the original query [36], or context-aware IR systems that integrate other well-known systems, such as Elastic Search [38].
7. **Improving Web Search and Recommendation:** Approaches that focus on boosting the capabilities of web search and recommendation systems. This includes many different approaches. One hybrid system provides an error-tolerant and location-aware query autocompletion system, such as in [40]. Another approach to optimizing autocompletion is presented in [68], where the authors utilize the recent BERT model to enhance automatic query expansion, thereby improving the overall performance of the document retrieval system. In [41], the authors propose a new solution to query reformulation using the popular firefly algorithm, which showed better results than traditional genetic algorithms or particle swarm optimization. The authors of [44] propose a new hybrid query autocompletion system that is time-sensitive and personalized.

4.2. Addressing information retrieval challenges

The optimization methods identified above also present several challenges that information retrieval systems face. One of the most significant challenges faced by IR systems is algorithmic bias. This often stems from biased training data, which may reflect unequal or unbalanced representation across different groups or topics. Additionally, techniques that rely on user feedback are particularly susceptible, as they can reinforce existing patterns and preferences, further enhancing bias over time. This often results in over-personalization and a reduction in content diversity. However, recent research efforts have begun to address these concerns by leveraging advanced models, such as ColBERT, to enhance fairness and diversity in retrieval results, as demonstrated in [12].

One significant challenge in information retrieval is determining the relevance of results. Users often input ambiguous or vague queries, making it difficult to discern their needs. Understanding the context of a query is crucial to providing relevant results, yet this task is inherently complex due to the nuances of human language. Another challenge is the sheer volume and variety of data. With the advent of big data, managing and retrieving relevant information from massive datasets has become increasingly difficult. Additionally, IR systems must handle heterogeneous data types and integrate them into a cohesive retrieval system. Retrieving media content differs from retrieving textual content in terms of approach, as demonstrated in [36].

Understanding user intent and personalizing results further complicates the retrieval process. Accurately interpreting what the user wants

is difficult, especially when queries are complex or poorly defined. Tailoring search results to individual users based on their preferences and previous interactions adds another layer of complexity that must be addressed through profiling [26]. While effective, such personalization techniques often rely on detailed user data, including search history, click behavior, and demographic information, which raises significant privacy concerns. These concerns include the collection, storage, and potential misuse of sensitive user information, as well as the profiling of users without their consent. The risk of unintended data exposure highlights the urgent need for transparent data practices and privacy-preserving approaches, such as federated learning or differential privacy.

Even if security were not a concern, overreliance on personalization can lead to retrieval results becoming too specific for the user, and the system never offering the user new information or documents. It will only result in documents specific to the user's initial intentions. Because of that, with personalization, we must also include diversification, which will widen the search area and, at the same time, offer the user particular documents that are related to him, but also documents that are not that related to the user himself but are related to the general field or area of search.

Content-based image retrieval systems present their challenges, the most significant of which is the semantic gap. The semantic gap presents a significant difference between high-level human concepts and low-level image features, as well as the gap between the query image and the images retrieved from the database [36]. There are several ways to address this gap. One is object ontology, which often utilizes tags and adds a semantic layer to various types of media content. Another approach is machine learning, where the system learns high-level concepts by analyzing various image features. Another approach is relevance feedback, which involves the user manually responding to the results returned by the IR system based on how well they match the initial query [36].

Real-time retrieval poses its own set of challenges. Ensuring that information retrieval is fast and efficient, particularly for time-sensitive queries, requires advanced algorithms and robust infrastructure. Handling dynamic data and providing up-to-date information is also critical in this context.

Another challenge arises when we attempt to retrieve information for documents written in different languages. Overcoming language barriers to retrieve relevant information across different languages and addressing translation issues requires sophisticated natural language processing (NLP) capabilities or highly specialized systems that cater to a specific language. This is illustrated in [30], which specifically addresses the Gujarati language.

4.3. Comparison of the analyzed studies

4.3.1. Strengths and weaknesses of optimization techniques

After analyzing the reviewed studies and the optimization techniques on which they are based, we will present a variety of strengths and weaknesses for these techniques in this section.

Query suggestion is a technique that boasts numerous strengths, making it one of the most popular optimization techniques. For one, it improves user experience by guiding users toward more refined and relevant queries. Additionally, through these suggestions, the system may also correct various spelling errors or typos that may appear. Context awareness is another notable strength of these techniques. It allows systems to leverage context from previous searches, user history, and current trends to provide highly relevant suggestions.

Conversely, users might rely too heavily on suggestions, potentially reducing the diversity of information they explore. Additionally, the previously mentioned context awareness can sometimes be a weakness if the system misinterprets the user's query context, which may lead to irrelevant suggestions. Additionally, these systems can become highly resource-intensive.

Relevance feedback is another prevalent optimization technique. There are many reasons, one being the degree of personalization they bring and increased accuracy. The system can tailor search results to each user by incorporating user feedback. User feedback also makes users feel more engaged with the search process, as they can directly influence the quality of the results they receive. Throughout this entire interaction, the system continuously evolves, resulting in improved search results with each subsequent interaction. This technique is also not ideal. Providing feedback can be time-consuming and may require significant effort from users, which not all users are willing to invest. Additionally, overreliance on feedback, particularly pseudo-relevance feedback, can result in feedback loops that repeatedly present users with similar, non-diverse content, thereby raising fairness concerns. Recent research has attempted to address these issues using advanced models such as ColBERT [5]. These systems can also show scaling issues. Processing and integrating feedback for large numbers of users can be a resource-intensive task.

Personalization is a technique often integrated with other optimization techniques. It ensures that the content presented to users is highly relevant to their interests, thus improving satisfaction. Websites typically strive to implement some form of personalization, as tailored content often results in higher interaction rates (such as clicks and likes). This comes with a cost, as personalization often requires collecting and analyzing personal data, which raises concerns about user privacy and data security. Additionally, heavy reliance on users' past behavior can lead to limited or no exposure to diverse perspectives.

On the other end of the spectrum, we have **diversification**. Diversification ensures that a broader range of user interests and preferences are covered. By providing diverse results, the system also reduces redundancies. Another essential characteristic of this technique is that it helps counteract the tendency of recommendation systems to reinforce existing biases, thereby providing a broader view that includes underrepresented or less popular items. This can sometimes backfire, and these systems can sometimes lead to a decrease in precision, as the system might include results less relevant to the user's intent. Additionally, diversification algorithms can sometimes require more computational resources and data to function effectively, which can impact system performance and scalability [47].

Image retrieval techniques are highly effective for visual search tasks. They enable users to search for images based on visual content or by using the images themselves. Additionally, systems like these have a wide range of domain applications, from e-commerce to medical. One of the most significant weaknesses or problems introduced by image retrieval is the semantic gap. The semantic gap represents a difference between low-level visual features and high-level concepts of an image, or, in another way, a disparity between how machines and users perceive images. This difference can manifest in various stages of the search process, such as during query formulation or retrieval. Additionally, image retrieval can be highly computationally expensive and introduces scalability issues, as efficiently handling and indexing large image databases can be challenging.

As for the optimization methods under "Information Retrieval Improvements" and "Web Search and Recommendation Enhancement", they present more unique methods that combine the previously mentioned optimization techniques with some new, original approaches. These methods generally provide enhanced accuracy and efficiency, a reduction in retrieval time, and the capability to handle large and diverse datasets [34]. Conversely, they can sometimes be significantly complex to implement or require high computational resources [32,48]. Additionally, they can often limit their use in other domains [32]. Another thing to note is that some of these approaches may exhibit scalability issues when handling large user databases.

4.3.2. Effectiveness comparison

The effectiveness of optimization techniques varies depending on the specific metrics used and the study context. However, some general trends can be observed:

1. **Query suggestion and relevance feedback:** These techniques often enhance precision and recall, resulting in improved retrieval of relevant results.
2. **Personalization:** This technique yields highly relevant results tailored to the user's interests and needs. Personalized approaches can lead to increased user satisfaction and engagement, as measured by click-through rates and time spent metrics.
3. **Diversification:** Diversification techniques can improve the overall coverage of search results, reducing the likelihood of missing relevant information. These techniques also cover a more comprehensive range of user interests and preferences, which can be particularly useful in scenarios where user intent is ambiguous. One more important characteristic is that diversification methods help counteract the tendency of recommendation systems to reinforce existing biases, providing a broader view that includes underrepresented or less popular items.
4. **Image retrieval:** These techniques provide enhanced search capabilities, enabling users to search for images based on visual content, such as colors, shapes, and patterns, thereby making it easier to find visually similar images. Additionally, in some cases, users can perform searches using images (e.g., uploading a photo) instead of relying solely on text, which can be more intuitive and user-friendly. Another strength of these systems is that they generally apply in many domains.

4.4. Final remarks

Table 2 summarizes the benefits and limitations of the different optimization techniques covered in the review.

5. Conclusion and future directions

5.1. Summary of key contributions

This systematic review presents an in-depth examination of optimization techniques in Information Retrieval, highlighting the fundamental shift toward AI-driven approaches that are redefining the field. We have synthesized the literature to categorize key strategies, ranging from foundational techniques such as query suggestion and relevance feedback, to adaptive user-centric systems based on personalization and diversification. Crucially, this review has integrated the latest advancements in deep learning, including the impact of transformer architectures and generative approaches like RAG. Furthermore, we have addressed the often-overlooked ethical dimension, discussing the challenges of algorithmic bias, fairness, and privacy, as well as proposed mitigation strategies. In doing so, this work offers a comprehensive view that connects technical advancements with the practical and societal implications for the field of decision analytics.

5.2. Limitations of the review

Despite their contributions, these optimization methods have several limitations. Data dependency presents one of those limitations. There are two ways in which this limitation might be presented. Personalization techniques can be seen as overreliance on user data, which might raise concerns about privacy and potential bias. Conversely, optimization methods that, at their core, incorporate a machine learning algorithm inherently rely heavily on training data.

Another one is computational complexity. Many higher-level, more sophisticated techniques can have higher computational costs, which can impact real-time performance. This is primarily evident in machine learning techniques and complex ranking algorithms. This is also prevalent in content-based image retrieval systems, which extract and compare images and their features, a task generally considered complex [36]. Similar behavior can be observed in the case of [34], where the authors proposed a new hybrid model dealing with complex document reranking. One more example can be found in [48], where the system integrates several components for query expansion, reranking, and indexing, which might require considerable computational resources and tuning.

One more limitation worth mentioning is the domain specificity. Certain optimization methods may be more effective in specific domains, thus limiting their general applicability. A simple example of this is image retrieval. Another example can be seen in [32], where the authors created a system that serves the purpose of retrieving information related to COVID-19, but unfortunately, it cannot be used to retrieve documents related to any other medical topic. Similarly, in [51], the authors proposed a new job-matching system. Still, it also has limitations that become apparent when adapting the system to different types of job markets or non-standard résumé formats.

Despite efforts to be comprehensive, this review has several limitations. First, the search strategy, although systematic, relied on a specific set of keywords and databases, which could have led to the exclusion of relevant studies using different terminology or indexed in other databases. Keyword selection bias is a common challenge in systematic reviews [25]. Second, the automated NLP filtering process, while efficient, may have missed relevant studies that did not match the predefined patterns. Third, the review was limited to English-language publications, which could introduce a linguistic bias and overlook important developments in other regions. Finally, the focus on peer-reviewed literature may have introduced publication bias, and the reviewers' perspectives inherently influence the interpretation of findings.

5.3. Avenues for future research

The field of IR optimization is poised for continued innovation. Based on the findings and identified gaps, we propose the following directions for future research:

- **Ethical AI by Design**: Future research must move beyond post-hoc bias mitigation. The focus should be on developing IR algorithms that are "ethical by design", incorporating fairness, transparency, and accountability as primary optimization objectives from the outset.
- **Efficiency and Scalability of AI Models**: As language models grow larger, research into knowledge distillation, quantization, and other model compression techniques will be crucial to enable the deployment of these powerful systems in resource-constrained, real-time production environments.
- **Generative IR and LLMs**: The RAG paradigm is still in its early stages of development. Future research will explore more sophisticated architectures that more tightly integrate the retrieval and generation processes, as well as the use of LLMs not only as answer generators but also as reasoning engines that can actively guide the search process [4].
- **Quantum IR Optimization**: As suggested by a reviewer, an emerging and promising frontier is the application of quantum computing to IR optimization problems. Quantum Annealing (QA) is a paradigm well-suited for solving complex combinatorial optimization problems that are common in IR [14].

Tasks such as feature selection, clustering, and ranking optimization can be formulated as Quadratic Unconstrained Binary Optimization (QUBO) problems and solved on quantum annealers [14]. While research in this area is still in its infancy, theoretical models suggest

Table 2
Optimization technique summary.

Optimization Technique	Benefits	Limitations
Query Suggestion	- Improves query quality, reduces irrelevant results - Helps in correcting spelling errors and typos - Can leverage context from previous searches or history	- May introduce biases - Potential over-reliance on suggestions - Context misinterpretation, which can lead to suggestions that are off the mark
Relevance Feedback	- Enhances relevance based on user interaction - Active participation of the user	- Relies on user feedback, may not capture all user preferences - Providing feedback requires time and effort from the user
Personalization	- Tailors results to user needs, improves satisfaction - Tailored content leads to higher interaction rates (clicks, likes, etc.)	- Requires extensive user data, may raise privacy concerns - Overuse leads to limited or no exposure to diverse perspectives
Diversification	- Provides diverse results, reduces redundancy - Reduced algorithmic bias	- May not always align with user preferences - Lower precision
Image Retrieval	- Effective for visual search tasks - Possibility for users to search with images - Possibility of application in various domains	- Limited by the semantic gap between visual features and textual queries - High computational requirements - Scalability issues due to large image databases
Information Retrieval Improvement	- Enhanced accuracy and efficiency - Capability to handle large datasets	- May require domain-specific adaptations - Can be a computationally challenging task
Web Search and Recommendation Enhancement	- Improves web search accuracy and recommendation relevance - Enhanced user experience	- May be affected by data quality and biases - Can be a computationally challenging task - Potential issues with scaling and real-life applicability

that quantum approaches could offer significant efficiency advantages for certain computationally intensive problems, thereby opening new avenues for advancement in the field [7].

In conclusion, this review not only consolidates state-of-the-art optimization strategies but also provides significant insights for researchers and practitioners to develop information retrieval systems further. The ramifications of these findings can inform future studies, helping to develop more efficient, equitable, and user-friendly information retrieval systems that support informed decision-making.

Funding statement

This work is funded by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações. It is also funded by FCT Pluriannual Funding UID/308: Instituto de Engenharia de Sistemas e Computadores de Coimbra - INESC Coimbra.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] J. Liu, et al., Data mining and information retrieval in the 21st century: A bibliographic review, Comput. Sci Rev 34 (2019) 100193, http://dx.doi.org/10.1016/j.cosrev.2019.100193.
[2] S. Ibrihich, A. Oussous, O. Ibrihich, M. Esghir, A review on recent research in information retrieval, Procedia Comput. Sci 201 (2022) 777–782, http://dx.doi.org/10.1016/j.procs.2022.03.106.
[3] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[4] C. Raffel, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.
[5] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020.
[6] B. Warner, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024, arXiv preprint arXiv:2412.13663.
[7] S. Hofstätter, A. Ram, S. Althammer, Lightning IR: A PyTorch lightning-based framework for information retrieval, 2024, arXiv preprint arXiv:2411.04677.
[8] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, Adv. Neural Inf. Process. Syst. (2020) 9459–9474.
[9] Amugongo Lameck Mbangula, et al., Retrieval augmented generation for large language models in healthcare: a systematic review, PLoS Digit. Health 4 (6) (2025) 1–33, http://dx.doi.org/10.1371/journal.pdig.0000877.
[10] Y. Ding, et al., A survey on RAG meets LLMs: Towards retrieval-augmented large language models, 2024, arXiv preprint arXiv:2405.06211.
[11] M.A.K. Raiaan, et al., A systematic review of hyperparameter optimization techniques in convolutional neural networks, Decis. Anal. J. 11 (2024) 100470, http://dx.doi.org/10.1016/j.dajour.2024.100470.
[12] M.O. Ayemowa, R. Ibrahim, Y.A. Bena, A systematic review of the literature on deep learning approaches for cross-domain recommender systems, Decis. Anal. J. 13 (2024) 100518, http://dx.doi.org/10.1016/j.dajour.2024.100518.
[13] Milvus, How do transformer models enhance IR? 2024.

[14] N. Keshav, A. Singh, Colbert-XM: A modular multi-vector representation model for zero-shot multilingual information retrieval, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025.

[15] PRISMA-P Group, et al., Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement, Syst Rev 4 (2015) 1, http://dx.doi.org/10.1186/2046-4053-4-1.

[16] Y. Gao, et al., Retrieval-augmented generation for large language models: A survey, 2023, arXiv preprint arXiv:2312.10997.

[17] B. Yadav, Retrieval-augmented generation (RAG): The definitive guide 2025, 2025.

[18] T. Matos, D. Vornicoglo, P.J. Coelho, E. Zdravevski, C. Albuquerque, I.M. Pires, Can sensors be used to measure the arm curl test results? a systematic review, Discov. Appl. Sci. 6 (2024) 48, http://dx.doi.org/10.1007/s42452-024-05643-5.

[19] C.L. Gabriel, et al., Mobile and wearable technologies for the analysis of ten meter walk test: A concise systematic review, Heliyon 9 (2023) e16599, http://dx.doi.org/10.1016/j.heliyon.2023.e16599.

[20] E. Zdravevski, et al., Automation in systematic, scoping and rapid reviews by an NLP toolkit: A case study in enhanced living environments, in: Enhanced Living Environments, Springer International Publishing, 2019, pp. 1–18, http://dx.doi.org/10.1007/978-3-030-10752-9_1.

[21] P. Shinde, R. Waghmode, D. Lokare, P. Halgaonkar, Efficient query suggestion system using users search history, in: 2018 International Conference on Information, Communication, Engineering and Technology, ICICET, IEEE, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICICET.2018.8533799.

[22] I. Badarinza, A. Sterca, D. Bufnea, V. Niculescu, Integration challenges for a web-based personalized query suggestions system in information retrieval, in: 2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications, SERA, IEEE, 2021, pp. 2–9, http://dx.doi.org/10.1109/SERA51205.2021.9509276.

[23] T.O. Kehinde, F.T.S. Chan, S.H. Chung, Scientometric Review and Analysis of Recent Approaches To Stock Market Forecasting: Two Decades Survey, Elsevier Ltd., 2023, http://dx.doi.org/10.1016/j.eswa.2022.119299.

[24] O.J. Adeleke, K.D. Jovanovich, S. Ogunbunmi, O. Samuel, T.O. Kehinde, Comprehensive exploration of smart cities: A systematic review of benefits, challenges, and future directions in telecommunications and urban development, IEEE Sensors Rev. 2 (2025) 228–245, http://dx.doi.org/10.1109/sr.2025.3569239.

[25] Y. Lu, Y. Li, M. Xu, W. Hu, A user model based ranking method of query results of meta-search engines, in: 2015 International Conference on Network and Information Systems for Computers, IEEE, 2015, pp. 426–430, http://dx.doi.org/10.1109/ICNISC.2015.123.

[26] H. Kumar, S. Lee, H.-G. Kim, Exploiting social bookmarking services to build clustered user interest profile for personalized search, Inf Sci (N Y) 281 (2014) 399–417, http://dx.doi.org/10.1016/j.ins.2014.05.008.

[27] K. Makvana, P. Shah, P. Shah, A novel approach to personalize web search through user profiling and query reformulation, in: 2014 International Conference on Data Mining and Intelligent Computing, ICDMIC, IEEE, 2014, pp. 1–10, http://dx.doi.org/10.1109/ICDMIC.2014.6954221.

[28] S. Preetha, K.N.V. Shankar, Personalized search engines on mining user preferences using clickthrough data, in: International Conference on Information Communication and Embedded Systems (ICICES2014), IEEE, 2014, pp. 1–6, http://dx.doi.org/10.1109/ICICES.2014.7033953.

[29] S. Yigit-Sert, I.S. Altingovde, C. Macdonald, I. Ounis, Ö. Ulusoy, Supervised approaches for explicit search result diversification, Inf Process. Manag. 57 (2020) 102356, http://dx.doi.org/10.1016/j.ipm.2020.102356.

[30] M. Servajean, R. Akbarinia, E. Pacitti, S. Amer-Yahia, Profile diversity for query processing using user recommendations, Inf. Syst. 48 (2015) 44–63, http://dx.doi.org/10.1016/j.is.2014.09.001.

[31] X. Tian, Y. Lu, N. Stender, L. Yang, D. Tao, Exploration of image search results quality assessment, IEEE Trans Big Data 1 (2015) 95–108, http://dx.doi.org/10.1109/TBDATA.2015.2497710.

[32] E. Assia, S.M. Abdelouahed, A. Abdellah, Image mining: Improving image retrieval technique using variational models and relevance feedback, in: 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET, IEEE, 2023, pp. 1–6, http://dx.doi.org/10.1109/IRASET57153.2023.10152921.

[33] P.L.K., S. Sharmila T., Dominant color and uniform local binary pattern based image retrieval, in: 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT, IEEE, 2021, pp. 1–5, http://dx.doi.org/10.1109/ICAECT49130.2021.9392616.

[34] A.P. Bhopale, A. Tiwari, Swarm optimized cluster based framework for information retrieval, Expert Syst. Appl. 154 (2020) 113441, http://dx.doi.org/10.1016/j.eswa.2020.113441.

[35] Y. Tang, H. Wang, K. Guo, Y. Xiao, T. Chi, Relevant feedback based accurate and intelligent retrieval on capturing user intention for personalized websites, IEEE Access 6 (2018) 24239–24248, http://dx.doi.org/10.1109/ACCESS.2018.2828081.

[36] S. Neji, T. Chenaina, A.M. Shoeb, L.B. Ayed, Hyra: An effective hybrid ranking model, Procedia Comput. Sci 192 (2021) 1111–1120, http://dx.doi.org/10.1016/j.procs.2021.08.114.

[37] A. Wen, Y. Wang, V.C. Kaggal, S. Liu, H. Liu, J. Fan, Enhancing clinical information retrieval through context-aware queries and indices, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 2800–2807, http://dx.doi.org/10.1109/BigData47090.2019.9006241.

[38] M. Aravind, S. Viswanath, N. Mohan, R. Adarsh, J. Bhaskar, A modified medical information retrieval system, in: 2019 IEEE 9th International Conference on Advanced Computing, IACC, IEEE, 2019, pp. 218–222, http://dx.doi.org/10.1109/IACC48062.2019.8971587.

[39] A. Khader, F. Ensan, Learning to rank query expansion terms for COVID-19 scholarly search, J Biomed Inf. 142 (2023) 104386, http://dx.doi.org/10.1016/j.jbi.2023.104386.

[40] J. Wang, C. Lin, Fast error-tolerant location-aware query autocompletion, in: 2020 IEEE 36th International Conference on Data Engineering, ICDE, IEEE, 2020, pp. 1998–2001, http://dx.doi.org/10.1109/ICDE48307.2020.00223.

[41] M. Zeboudj, K. Belkadi, Web query reformulation using FireFly algorithm, in: 2020 Second International Conference on Embedded & Distributed Systems (EDiS), IEEE, 2020, pp. 87–90, http://dx.doi.org/10.1109/EDiS49545.2020.9296463.

[42] C.L. Smith, J. Gwizdka, H. Feild, The use of query auto-completion over the course of search sessions with multifaceted information needs, Inf Process. Manag. 53 (2017) 1139–1155, http://dx.doi.org/10.1016/j.ipm.2017.05.001.

[43] K.C. Jena, S. Mishra, S. Sahoo, B.K. Mishra, Principles, techniques and evaluation of recommendation systems, in: 2017 International Conference on Inventive Systems and Control, ICISC, IEEE, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICISC.2017.8068649.

[44] F. Cai, S. Liang, M. De Rijke, Prefix-adaptive and time-sensitive personalized query auto completion, IEEE Trans. Knowl. Data Eng. 28 (2016) 2452–2466, http://dx.doi.org/10.1109/TKDE.2016.2568179.

[45] S. Malik, M. Saleem, Interest indicators in structured scientific articles, Procedia Comput. Sci 116 (2017) 158–165, http://dx.doi.org/10.1016/j.procs.2017.10.063.

[46] V. Pouli, S. Kafetzoglou, E.E. Tsiropoulou, A. Dimitriou, S. Papavassiliou, Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience, in: 2015 13th International Conference on Telecommunications (ConTEL), IEEE, 2015, pp. 1–8, http://dx.doi.org/10.1109/ConTEL.2015.7231205.

[47] I. Badarinza, A. Sterca, D. Bufnea, A dataset for evaluating query suggestion algorithms in information retrieval, in: 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), IEEE, 2019, pp. 1–6.

[48] V. Singh, A. Singh, Learn-as-you-go: Feedback-driven result ranking and query refinement for interactive data exploration, Procedia Comput. Sci 125 (2018) 550–559, http://dx.doi.org/10.1016/j.procs.2017.12.071.

[49] V. Muralikrishnan, B. Janakiraman, Firefly based optimization in web page recommendation system, in: 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), IEEE, 2018, pp. 96–101, http://dx.doi.org/10.1109/IC3IoT.2018.8668189.

[50] S. Guo, F. Alamudun, T. Hammond, RésuMatcher: A personalized résumé-job matching system, Expert Syst. Appl. 60 (2016) 169–182, http://dx.doi.org/10.1016/j.eswa.2016.04.013.

[51] B. Shah, J. Pareek, Query optimization for information retrieval in multilingual environment for E-governance resources, in: 2016 International Conference on ICT in Business Industry & Government, ICTBIG, IEEE, 2016, pp. 1–4, http://dx.doi.org/10.1109/ICTBIG.2016.7892650.

[52] Y. Cao, J. Fan, G. Li, A user-friendly patent search paradigm, IEEE Trans. Knowl. Data Eng. 25 (2013) 1439–1443, http://dx.doi.org/10.1109/TKDE.2012.63.

[53] D. Tranfield, D. Denyer, P. Smart, Towards a methodology for developing evidence-informed management knowledge by means of systematic review, Br. J. Manag. 14 (3) (2003) 207–222, http://dx.doi.org/10.1111/1467-8551.00375, SUBPAGE:STRING:ABSTRACT;WEBSITE:WEBSITE:PERICLES;REQUESTEDJOURNAL:JOURNAL:14678551;ISSUE:ISSUE:DOI.

[54] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W.M. Lim, How to conduct a bibliometric analysis: An overview and guidelines, J Bus Res 133 (2021) 285–296, http://dx.doi.org/10.1016/J.JBUSRES.2021.04.070.

[55] Y. Xiao, M. Watson, Guidance on conducting a systematic literature review, J Plan Educ Res 39 (1) (2019) 93–112, http://dx.doi.org/10.1177/0739456X17723971, PAGE:STRING:ARTICLE/CHAPTER.

[56] M.F. Hasani, R. Mandala, Improving pseudo relevance feedback with term relationship using firefly algorithm, in: 2020 IEEE International Conference on Sustainable Engineering and Creative Computing, ICSECC, IEEE, 2020, pp. 146–151, http://dx.doi.org/10.1109/ICSECC51444.2020.9557560.

[57] R. Nogueira, J. Lin, S. Lin, Document ranking with a pre-trained sequence-to-sequence model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4799–4803.

[58] S. Lin, J. Hilton, O. Evans, A survey of transformers, 2021, arXiv preprint arXiv:2106.04554.

[59] Z. Xu, G. Raskutti, A survey of model architectures in information retrieval, in: To Appear in Proceedings of the ACL Rolling Review, 2025.

[60] R. Nogueira, J. Lin, From zero to hero: A simple pre-training approach for document ranking, 2019, arXiv preprint arXiv:1911.05969.

[61] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, ColBERTv2: Effective and efficient retrieval via lightweight late interaction, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022.

[62] S.A. Lee, A. Wu, J.N. Chiang, Clinical ModernBERT: An efficient and long context encoder for biomedical text, 2025, [Online]. Available: https://arxiv.org/abs/2504.03964.

[63] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2019.

[64] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[65] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[66] A. Asai, J. Kasai, H. Hajishirzi, L. Zettlemoyer, A. Cohan, Retrieval-based language models and applications, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts), 2023.

[67] Y. Huang, J. Huang, A survey on retrieval-augmented text generation for large language models, 2024, arXiv preprint arXiv:2404.10981.

[68] D. Vishwakarma, S. Kumar, Fine-tuned BERT algorithm-based automatic query expansion for enhancing document retrieval system, Cogn. Comput. 17 (2025) 23, http://dx.doi.org/10.1007/s12559-024-10354-5.