

RESEARCH ARTICLE

JASIST WILEY

Incremental refinement of relevance rankings: Balancing relevance depth and scope

Müge Akbulut¹  | Yaşar Tonta² 

¹Department of Information Management, Ankara Yıldırım Beyazıt University, Türkiye

²Department of Information Management, Hacettepe University, Türkiye

Correspondence

Müge Akbulut, Department of Information Management, Ankara Yıldırım Beyazıt University, Ankara, Türkiye.

Email: mugeakbulut@aybu.edu.tr; mugeakbulut@gmail.com

Funding information

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, National Postdoctoral Research Fellowship Program

Abstract

Delivering both relevant and topically diverse results is a key challenge in information retrieval (IR). This study introduces a hybrid method that incrementally refines rankings by combining probabilistic topic modeling (latent dirichlet allocation [LDA]) with citation-based pennant retrieval grounded in Relevance Theory (RT), optimizing for both relevance and diversity. A unified scoring function fuses topic and citation cues, while a tunable weight parameter makes the relevance–diversity trade-off explicit and transparent. Applying this RT-guided approach to the iSearch corpus (435,000 physics papers, 3.7 million citations) across 65 queries, we evaluate relevance and diversity via degree centrality, arXiv subject categories, and maximal marginal relevance (MMR). Unlike prior methods that apply LDA or pennant retrieval separately—or depend on post-hoc diversification such as MMR—this approach integrates them directly and efficiently: LDA runs on local multi-core CPUs, and pennant retrieval operates using sparse co-citation counts. The combined technique retrieves both central and peripheral works, producing a boundary-spanning effect that broadens conceptual coverage while improving normalized discounted cumulative gain (NDCG) performance. Interactive visualizations via LuminaCite depict how researchers can fine-tune results in real time to reveal latent or interdisciplinary connections. Sensitivity and robustness tests indicate that the proposed model is robust to hyperparameter variation, first-stage retriever choice, and query reformulation.

1 | INTRODUCTION

Information retrieval (IR) systems match user queries with relevant papers, typically by aligning query terms with the vocabulary used to describe content (Chimah & Ude, 2020). Yet this seemingly straightforward task is complicated by term mismatches, including synonymy, polysemy, and the inherent subjectivity and variability of language across users, time, and

context (Crestani, 2000; Ganguly & Jones, 2018; Swanson, 1986a). Relevance in this sense is not solely determined by lexical similarity but is shaped by users' cognitive states, expectations, and effort (Mizzaro, 1998; Saracevic, 2007). Because of the complex and dynamic nature of relevance, the ideal IR system retrieving all and only relevant papers remains a theoretical construct (Mizzaro, 1997). In practice, relevance rankings are indispensable tools that help users find

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

satisfactory results efficiently, even when perfect retrieval is unattainable (Lei et al., 2001; Tonta, 1992).

While IR systems typically rank papers by the degree of term-matching or by their estimated probability of relevance (Robertson, 1977), users often need more than just highly similar or redundant results. Exploratory tasks particularly benefit from diverse perspectives, varying subtopics, and alternative applications of the queried topic. Diversity, therefore, emerges as an essential success criterion alongside relevance for literature reviews, interdisciplinary research, and creative problem solving (Küçüktunç & Ferhatosmanoğlu, 2011; Poetz & Prügl, 2010). An effective IR system should retrieve topically relevant papers and encourage boundary spanning, cognitive shifts, and serendipitous discovery (Harter, 1992).

Topic modeling became a widely used technique for generating relevance rankings by uncovering latent thematic structures in large collections. Among these, Latent Dirichlet Allocation (LDA) is a prominent probabilistic model that infers topic-word and document-topic distributions from word frequency patterns (Blei et al., 2003; Li & McCallum, 2006). However, its reliance on the “bag-of-words” assumption limits its capacity to capture deeper semantic relationships, leading to terminological ambiguities and topic overlaps (Guo et al., 2016; Năstasă et al., 2024; Wu et al., 2007, 2008).

Citation-based methods provide an alternative path to relevance estimation. Co-citation patterns offer insights into conceptual proximity between papers and increase IR performance significantly (Mayr & Mutschke, 2013). They also allow for the detection of research communities and topical interest structures (Carevic & Schaer, 2014; Han, 2020; Knoth et al., 2017; Küçüktunç et al., 2015; White, 2010). Citation-enhanced retrieval, including literature-based discovery (Swanson, 1986b) and pattern-based relationship extraction, reveals indirect but meaningful connections across domains (Yang et al., 2017). However, citation-based approaches depend on citation accrual over time (“citation lag”) and often require a predefined seed paper (Ke et al., 2015). The search performance of the citation-based clusters can vary unpredictably (Bascur et al., 2023).

Word- and citation-based methods suffer from domain-specific vocabulary issues, citation delays, or dependence on predefined seed papers. For example, LDA may overlook synonymous or polysemous terms used across disciplines while citation-based systems may miss relevant, recent, or uncited papers (Küçüktunç et al., 2015; Zarrinkalam & Kahani, 2012). Nevertheless, they can complement one another. Combining citation data with topic modeling and textual approaches improves recall, retrieves influential works, and streamlines the performance of seed-based IR systems (Huang et al., 2016; Huang et al., 2018; Li et al., 2020;

Nguyen & Do, 2018; Sjögarde & Ahlgren, 2024; Wang et al., 2013; Xia et al., 2012; Zhou et al., 2017). Since different algorithms generate distinct sets of relevant papers, their integration promises richer, more complete retrieval, recognizing the complex interplay between textual topics and citation structures (Ballester & Penner, 2022; Croft, 2002).

This study proposes a hybrid retrieval approach integrating LDA topic modeling with citation-based pennant retrieval to produce rankings that are both topically relevant and diverse. The goal is to generate rankings that include both core and marginal papers that broaden the scope of the query, thereby achieving incremental refinement and boundary spanning. Pennant diagrams building on Relevance Theory (RT) and co-citation patterns provide contextual depth that complements the probabilistic output of LDA. The resulting system allows users to customize the relevance–diversity balance to suit their specific research needs.

Recent advances in Large Language Models (LLMs) offer alternative, semantic-rich approaches to retrieval (Brown et al., 2020; Vaswani et al., 2017), but their opaque architectures limit transparency and control, especially in structured tasks like literature reviews. By contrast, the proposed method provides an interpretable and user-controllable mechanism for exploratory and analytical search.

To assess its performance, we implemented the method on the iSearch dataset. Results demonstrate that augmenting LDA-based rankings with pennant retrieval improves both relevance and diversity, offering a transparent and adaptable solution for advanced academic search.

The main contribution of this study is that it integrates a topic model with White's pennant retrieval (2007a, 2007b) and Sperber and Wilson's (1995) RT. This integration uses a simple user-controlled weight that makes the balance between relevance and diversity explicit and is compatible with different seed selectors. We present an interpretable RT-guided integration of pennant retrieval citation signals and LDA topic cues, and release the LuminaCite prototype and companion scripts to support transparency and reuse. This study represents the first “large-scale” performance test of pennant retrieval with expert relevance assessments based on the iSearch corpus. Previous research on pennant retrieval has primarily been exploratory and focused on case studies (Akbulut, 2016; Akbulut et al., 2020; White, 2007a, 2007b, 2009, 2010, 2015).

2 | LITERATURE REVIEW

The quest for relevance remains the central notion in IR (Lewandowski, 2009; Saracevic, 2007). Since the 1950s,

relevance-based ranking and evaluation mechanisms have underpinned the assessment of retrieval effectiveness (Saracevic, 2007; Verma et al., 2016). Traditional IR models emphasize topical relevance by exploiting textual information (Carevic & Schaer, 2014; White, 2007b), often neglecting novelty and diversity (Cosijn & Ingwersen, 2000). Semantic gaps that arise when query and document terms do not align demand retrieval strategies that go beyond lexical matching. Approaches addressing this include Natural Language Processing (NLP) techniques broadly falling into word- and citation-based categories (Cambria & White, 2014; Chen & Décary, 2018). RT serves as a conceptual framework for understanding how these methods operate and intersect.

2.1 | Word-based approaches

Probabilistic topic models uncover hidden thematic structures in large corpora by analyzing word co-occurrence patterns (Boyd-Graber & Blei, 2010). LDA, a widely adopted model, represents papers as mixtures of latent topics, each characterized by a probability distribution over words (Blei et al., 2003). This enables the extraction of thematic information beyond binary keyword matching (Grün & Hornik, 2011).

However, LDA's reliance on frequency-based and context-independent assumptions requires large corpora to produce coherent topics (Chen & Liu, 2014; Leydesdorff & Nerghes, 2017; Nguyen & Do, 2018; Xie et al., 2019), while paradoxically, increasing the number of topics may reduce coherence (Hecking & Leydesdorff, 2018). Term-level computation increases costs for large or multilingual corpora. Its "bag-of-words" assumption disregards word order and inter-document semantic relationships, limiting its ability to model contextual richness (Chang et al., 2009; Ekinci & İlhan Omurca, 2020). The model also struggles with synonymy and polysemy, thus introducing ambiguity. These limitations can result in relevant papers being overlooked and unrelated ones retrieved (Küçüktunç et al., 2015). Moreover, topic stability may be compromised due to the algorithm's sensitivity to input order (Agrawal et al., 2018). Attempts to mitigate these shortcomings include hybridizing LDA with citation data, using semi-supervised learning, or integrating context-aware models like BERT (Liu & Gu, 2023). Yet, fundamental semantic limitations persist.

2.2 | Citation-based approaches

Citation analysis compensates for the contextual limitations of word-based methods by revealing intellectual structure, thematic relationships, and scholarly influence.

Citations help map the conceptual landscape of scientific fields. Incorporating citation information into IR algorithms enhances ranking performance and supports recommender systems (Beel et al., 2016; Beel & Gipp, 2009; Pao, 1993).

Two foundational citation methods are bibliographic coupling (BC) and co-citation analysis. BC links papers sharing references, identifying direct, static, and backward-looking relationships (Akbulut et al., 2020; Carevic & Mayr, 2014; Kessler, 1963; Vergoulis et al., 2019). Yet, it fails to identify conceptually related works when few references overlap (Ma et al., 2022; Martinho, 2022; Nandy et al., 2024; Shen et al., 2019). In contrast, co-citation analysis captures dynamic, direct, and indirect connections among disciplines by measuring how often two papers are cited together and maps the evolving intellectual structure of fields effectively (Akbulut et al., 2020; Beel et al., 2016; Small, 1973; White & McCain, 1998). However, co-citation methods require time to accumulate data and can underperform in rapidly changing or under-cited fields. Topic-dependent variability also influences clustering effectiveness: fields like medicine and psychology are well represented, while natural sciences and geography often are not (Bascur et al., 2023, 2025).

To address these limitations, RT offers a cognitively grounded approach. Sperber and Wilson (1995) define relevance as a trade-off between cognitive effect and the processing effort. Higher relevance arises when significant cognitive effects (e.g., insight or contradiction resolution) are achieved with minimal mental effort. Howard D. White operationalized this theory in IR through "pennant retrieval" (White, 2007a, 2007b, 2009, 2010, 2015, 2016, 2018).

Pennant retrieval reinterprets the $tf \cdot idf$ formula from Salton et al. (1975), replacing tf (term frequency) with co-citation counts and idf (inverse document frequency) with total citation counts. In pennant retrieval, tf represents cognitive effect or topical relevance of a seed paper, and idf is a measure of processing effort or visibility. Thus, a highly co-cited paper is considered cognitively impactful whereas its total citation count represents ease of processing. This presents a more holistic understanding of relevance and yields a ranking that reflects both topical similarity and cognitive utility.

The pennant approach also helps track conceptual influence over time and identify interdisciplinary connections. For example, pennant retrieval outperformed the Web of Science "related records" feature for Maron and Kuhns' (1960) seminal paper, despite its sparse citation profile (Akbulut et al., 2020). However, limitations persist: it relies on seed papers, suffers from citation lag, and may not independently capture diversity (Table 1 compares these methods). These constraints highlight the need for hybrid systems.

TABLE 1 Comparison of basic citation analysis methods.

Feature	Co-citation analysis	Bibliographic coupling	Pennant retrieval (uses co-citations)
Definition	The frequency with which two papers are cited together	Two papers citing a common third paper	Ranks papers co-cited with a seed paper based on co-citation and total citation data
Unit of analysis	Cited papers	Citing papers	Papers co-cited with a seed article
Nature	Dynamic/forward-looking	Static/backward-looking	Dynamic/forward-looking
Key strength	Reveals evolving intellectual structures of research areas	Identifies historically related research	Integrates Relevance Theory, emphasizes cognitive effect and accessibility/diversity
Key limitation	Requires co-citations to accumulate	Static view may miss newer links	Requires a seed paper; dependent on existing citation data
Integration with Relevance Theory	No	No	Yes (core principle)

2.3 | Incremental refinement

High topical relevance alone is insufficient; diversity is equally critical for exploratory tasks. Achieving both simultaneously is challenging but essential, as diverse results improve discovery, enhance user experience, and engender serendipitous insights (Adomavicius & Kwon, 2011; Bradley & Smyth, 2001; Herlocker et al., 2004; McNee et al., 2006). Interdisciplinary results foster broader understanding (Abramo et al., 2018; Akbulut, 2016; Rafols, Leydesdorff, et al., 2012; Rafols, Porter, & Leydesdorff, 2012). The “filter bubble” effect of conventional recommender systems can reinforce siloed reading and citation habits by prioritizing items like past recommendations, potentially hindering the cross-disciplinary knowledge transfer essential for breakthrough discoveries (Cunningham et al., 2025).

Relevance ranking algorithms typically prioritize a list sorted by decreasing topical relevance, whereas diversification strategies aim for a balanced list encompassing a broader spectrum of topics or perspectives (Li et al., 2020). Techniques like maximal marginal relevance (MMR) address this need by balancing query relevance with novelty relative to already retrieved items (Carbonell & Goldstein, 1998). Similarly, re-ranking methods adjust topically ranked lists using factors such as diversity, popularity, or recency. Hybridization techniques like data fusion and cascade computation combine complementary strengths of different methods, often outperforming single source models (Baeza-Yates & Ribeiro-Neto, 1999; Jin et al., 2008; Ma et al., 2022; Meng et al., 2002). Citation-based clusters can effectively complement query-based retrieval by capturing relevant but lexically dissimilar papers (Bascur et al., 2023).

2.4 | State-of-the-art context and study scope

Recent advances in IR are driven by LLMs and neural retrievers, offering strong performance in semantic retrieval, contextual understanding, and answer generation (Brown et al., 2020; Devlin et al., 2019; Metzler et al., 2021; Vaswani et al., 2017). Retrieval-Augmented Generation (RAG) architectures can further extend them by incorporating external sources (Lewis et al., 2020). They are often computationally demanding and less interpretable than classical retrieval pipelines. Beyond computational cost and interpretability, a more fundamental architectural distinction is also relevant: using LLMs primarily for front-end query rewriting or expansion does not in itself address how retrieved papers are ranked. Although automatic query expansion improves recall, added terms introduce noise when they drift from the user's information need, thereby weakening ranking quality (Carpineto & Romano, 2012). Recent empirical evidence from AI-assisted discovery environments likewise suggests that such systems do not always provide a clear topical relevance advantage over conventional search results (Galbreath et al., 2025). By contrast, successful systems such as PubMed Best Match adopt a two-stage architecture combining first-stage retrieval with machine-learning-based re-ranking to improve result ordering (Fiorini et al., 2018).

Accordingly, rather than altering the query, this study focuses on incrementally refining an initial LDA-based ranking through transparent, traceable, and user-controllable pennant-based post-processing. Our LDA pipeline is locally computable and runs efficiently on multi-core CPUs, while pennant retrieval operates over co-citation counts. By contrast, LLM-based seed selection

or re-ranking introduces additional latency and hardware costs.

Moreover, beyond these architectural considerations, LLMs' role in IR remains uneven. They often function as AI-assisted interfaces layered over retrieval outputs rather than as transparent retrieval mechanisms (Urs, 2025). Research shows that LLMs struggle with complex queries, that performance varies across models and domains, that their longitudinal effect on IR benchmark studies is inconclusive, and that incomplete and fabricated answers remain a persistent concern (Bogers et al., 2025; Ravuri et al., 2026; Staudinger et al., 2026). Such fabricated outputs are linked in part to prevailing evaluation practices penalizing uncertainty and encouraging models to produce plausible but false statements rather than explicitly acknowledging uncertainty (Kalai et al., 2025).

Nonetheless, LLMs reshape IR at multiple levels, including retrieval, user interaction, and the broader evaluation ecosystem (Breuer et al., 2025). Here, we focus on retrieval and examine whether LDA topic cues can be incrementally refined with citation signals from pennant retrieval in a transparent and user-controlled way. This design remains compatible with LLM-guided seed selection while preserving interpretability for literature reviews.

We also acknowledge the growing use of LLM-assisted evaluation and its associated risks. Language models assign relevance more liberally than human assessors, are more likely to mark nonrelevant papers containing query terms as relevant, and remain vulnerable to spamming, bias, and hallucination effects (Alaofi et al., 2024; Bogers et al., 2025). They cannot reliably replace human judgment in automated relevance assessment or seed selection workflows (Breuer et al., 2025; Faggioli et al., 2024; Soboroff, 2025). Therefore, the present study relies exclusively on established IR metrics.

This study does not aim to benchmark against current end-to-end LLM-based systems, given the substantial differences in dataset scale, task formulation, and computational requirements. Instead, it addresses a narrower but well-defined problem: improving LDA-based retrieval through pennant-based post-processing, grounded in RT, to balance topical relevance and diversity in academic search. Within this scope, we retained LDA as the main first-stage retriever because it is interpretable, replicable, and computationally efficient, while also providing a transparent baseline for isolating the contribution of the pennant-based post-processing step.

2.5 | Positioning the study and its contribution

While topic models capture latent thematic structure, they often provide limited contextual and relational

nuance. Pennant retrieval, by contrast, is effective at highlighting influential and conceptually related works, but it depends on seed papers and existing citation patterns. Although prior studies have combined topic modeling with bibliometric signals in various ways (e.g., Nguyen & Do, 2018; Zhou et al., 2017; Zou et al., 2021), few integrated it with RT (e.g., Akbulut, 2022; Akbulut & Tonta, 2022).

This study addresses that gap by proposing and evaluating a framework that incrementally refines LDA-based rankings through context and influence signals derived from White's pennant retrieval. It supports both topical relevance and interpretable, user-controllable diversity (Caballero & Akella, 2012; Jiang et al., 2014). Within this framework, LDA contributes thematic breadth, corresponding to cognitive effect across topics, whereas pennant retrieval contributes citation-based coherence by reducing processing effort around seed papers. We make this trade-off explicit through transparent normalization and weighting, making the integration process readily interpretable to users.

This study does not introduce a new topic model or a new pennant retrieval variant but develops an interpretable integration framework that makes the relevance-diversity trade-off explicit, traceable, and adjustable. It offers a transparent post-processing design for academic retrieval to uncover conceptually related papers beyond disciplinary boundaries, thereby supporting both relevance and diversity in retrieval results. Additionally, it presents LuminaCite, an interactive prototype platform that enables user-driven query refinement and serves as a proof of concept for the proposed method, demonstrating its practical value for user-centered literature discovery.

This hybrid approach supports improved relevance rankings and the discovery of marginal and interdisciplinary works. In doing so, it addresses common shortcomings of standard IR systems, such as repetitive result sets, echo chambers, and filter bubbles (Cunningham et al., 2025). By combining thematic extraction with citation-based diversity through the LuminaCite platform, this study lays the groundwork for future discovery-oriented and user-controllable academic retrieval systems.

3 | DATA SOURCES AND METHODS

3.1 | Data sources

This study used the iSearch test collection comprising some 435,000 physics papers from arXiv and over 3.7 million citations (Lykke et al., 2010). It also includes 65 search queries, each embedded within a detailed scenario ("topic") simulating a realistic information-seeking situation providing keywords, describing information need, background knowledge,

and an ideal answer. The “search_terms” provided in each scenario were used as the operational query. Crucially, the collection is enriched with expert relevance assessments for 65 queries, thereby offering a robust basis for evaluation. Missing metadata such as subject categories was retrieved from [arXiv.org](https://arxiv.org). Papers are distributed unevenly across physics subcategories, “astrophysics” comprising 22% of all papers. Some subject categories are listed under multiple archives.

The iSearch citation network is sparse (density = 0.0021), with a median of five citations per paper (mean = 15). Co-citation counts are correspondingly low (median = 1; mean = 2), constraining the performance of the pennant retrieval algorithm.

3.2 | Methods

This research employed both word frequency-based approaches and citation relationships to generate relevance rankings. Relevance was measured using centrality values while diversity was assessed based on the primary subject categories assigned in arXiv. Both intellectual content and document context were considered. The process creates an integrated list (LDA + pennant retrieval) and personalizes the rankings (re-ranking) (Figure 1).

3.2.1 | Topic modeling LDA

LDA is a hierarchical Bayesian model assuming that each paper is a probabilistic mixture of topics, and each word is

associated with them, enabling the estimation of a paper's relevance to a specific topic as a probabilistic value.

Formula 1 expresses the underlying generative process of the LDA algorithm and the joint probability of the observed data: papers (p), words (W), topics (Z). Each paper has a topic mixture (θ , controlled by the parameter α) determining the proportion of topics it comprises. Each topic, in turn, has a unique word distribution (φ , controlled by the parameter β) indicating which words it likely contains. Each word in a paper is generated by first probabilistically selecting one of the associated topics and then probabilistically drawing a word from that topic's distribution. The overall probability of the model is calculated by multiplying the probabilities of the formation of paper-topic, topic-word, and word selection distributions (Formula 1). In essence, it forms the mathematical basis for probabilistically modeling the latent topic structure of texts based on observed word data.

$$p(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^M p(\theta_j; \alpha) \prod_{i=1}^K p(\varphi_i; \beta) \prod_{t=1}^N p(Z_{j,t} | \theta_j) p(W_{j,t} | \varphi_{Z_{j,t}}) \quad (1)$$

The success of this model largely depends on data quality; therefore, a careful preprocessing and data cleaning process was conducted before applying the LDA algorithm to the iSearch dataset. It comprised lowercasing, stemming, and tokenizing words in the titles and abstracts of papers and excluding stop words for physics (ADS Team, 2008). After removing 3617 papers (0.8%) with fewer than 25 words in their metadata, the LDA algorithm was applied to the remaining 431,196 papers using the Gensim (version 4.1.2) library in Python (settings: alpha = “symmetric”, iterations = 50, gamma_threshold = 0.001, minimum_probability = 0.01).

Four metrics (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths & Steyvers, 2004) were used to determine the optimum total number of topics (k), a critical parameter for LDA (Gläser et al., 2017). It ranged between 110 and 130 for the iSearch dataset (see https://mugeakbulut.com/incremental_refinement/optimal_topic_number_for_iSearch.html). One (Deveaud et al., 2014) was excluded due to inconsistencies with the others as it was developed for different evaluation contexts (Holliger, 2018), and k was set at 130. The effect of Dirichlet prior parameters, α and β , on prediction performance is limited in large corpora (Wallach et al., 2009; Zhang et al., 2015; Zhang et al., 2016). Therefore, we adopted a symmetric document-topic prior and standard convergence settings, consistent with common practice for large corpora. The statistics (e.g., unique word count, paper count) calculated before applying the LDA algorithm to iSearch dataset are compatible with a medium-sized model.

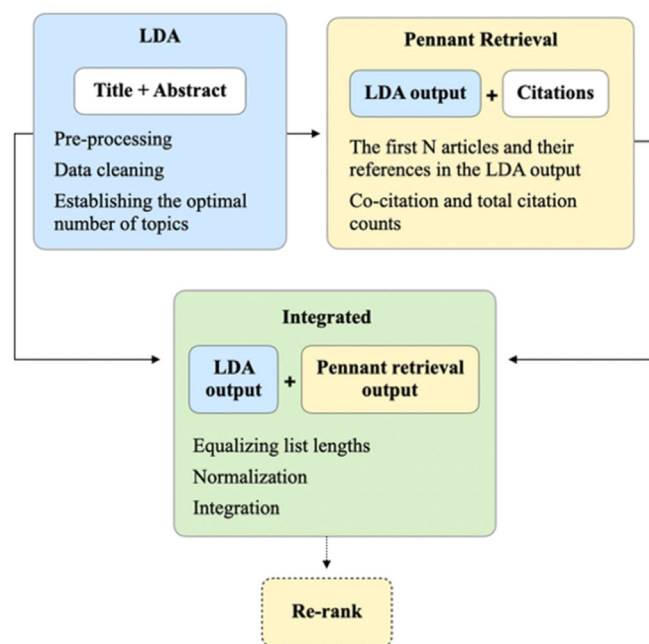


FIGURE 1 Relevance ranking process.

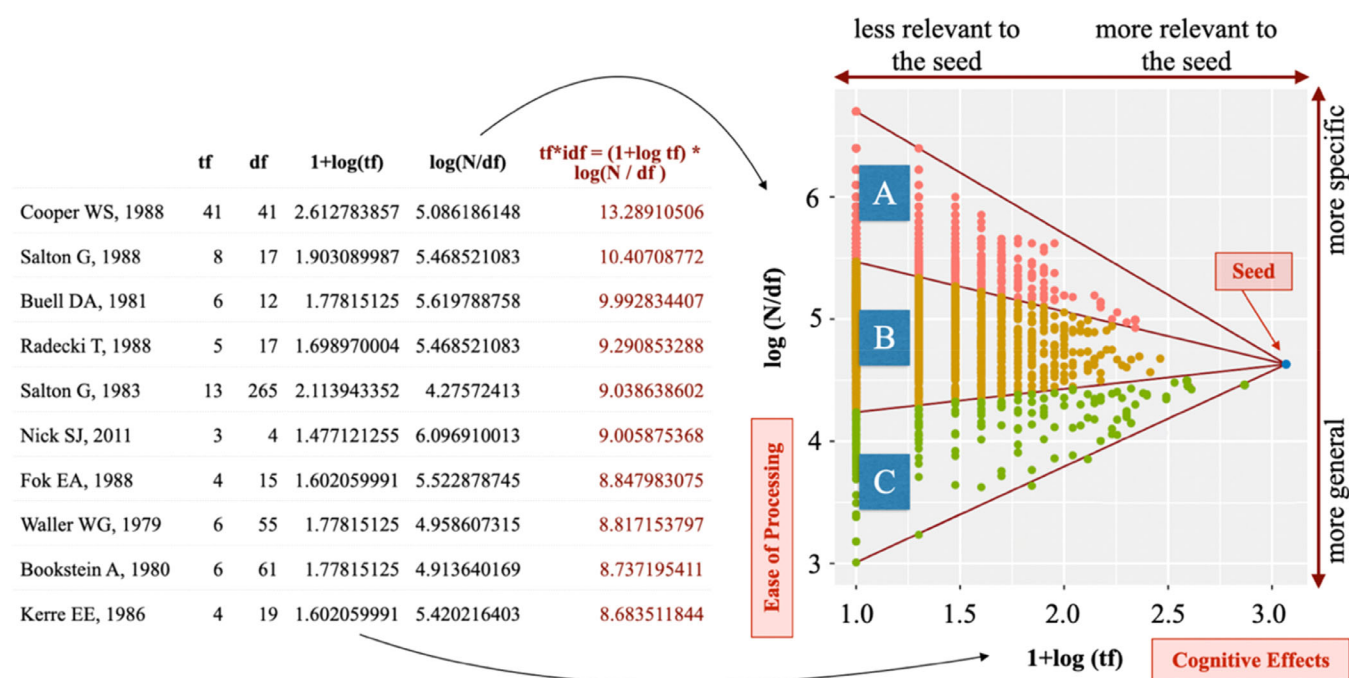


FIGURE 2 Example pennant retrieval visualization.

3.2.2 | Pennant retrieval algorithm

This method calculates relevance based on co-citation frequency using a modified $tf*idf$ weighting formula (Manning & Schütze, 2000):

$$\text{Relevance} = tf * idf = (1 + \log(tf)) * \log(N/df) \quad (2)$$

Here, tf represents the number of co-citations a paper shares with the seed paper(s), df denotes the total citation count of the co-cited paper, and N is the overall size of the iSearch dataset ($N = 434,813$ papers).

High relevance scores are obtained when tf and idf values are similar (Akbulut, 2016). Pennant diagrams visualize these relationships by plotting cognitive effect ($1 + \log(tf)$) against ease of processing ($\log(N/df)$). Figure 2 shows an example for Cooper (1988) as seed. Papers that appear in section A tend to be semantically closer to the seed in terms of relevance and specificity. The ones in sections B and C are decreasingly less relevant, thereby making them harder to associate with the seed.

3.2.3 | Seed selection strategy (modular design)

Pennant retrieval is effective for identifying core sources, but requires seed paper(s), and ranking cannot be fixed in advance because it depends on observed co-citation counts. We mitigated these limitations by using pennant

retrieval as a post-processing step over an initial ranked list: all papers in the collection remained available in the final integrated ranking, while the influence of pennant retrieval was concentrated on top of the list.

For seed selection, we identified the top five query-relevant papers as seeds via LDA and applied pennant retrieval to them (Figure 3). When seed selection is noisy, post-processing based on co-citation can amplify that noise within the local citation neighborhood. This risk is especially relevant in sparse citation settings. Therefore, we chose LDA-based seeds, as they provided a transparent, replicable, interpretable, and stable starting point to evaluate pennant retrieval's contribution.

This modular framework is “seed-agnostic”. The user-controlled downstream integration procedure of min-max normalization remains unchanged regardless of the first-stage retriever and is compatible with LDA and dense retrievers of SBERT and SPECTER, as examined later. We set the number of seeds to five to balance stability and computational cost.

The platform operates over the iSearch collection. It combines LDA, applied to titles and abstracts, with citation-weighted ranking based on the pennant retrieval algorithm. Query processing is accelerated through Pandas-based data handling and parallel computation on a macOS ARM64 system with a 10-core CPU. The platform is implemented in Python 3.9.12 using the Flask framework and an SQLite backend for efficient retrieval management.

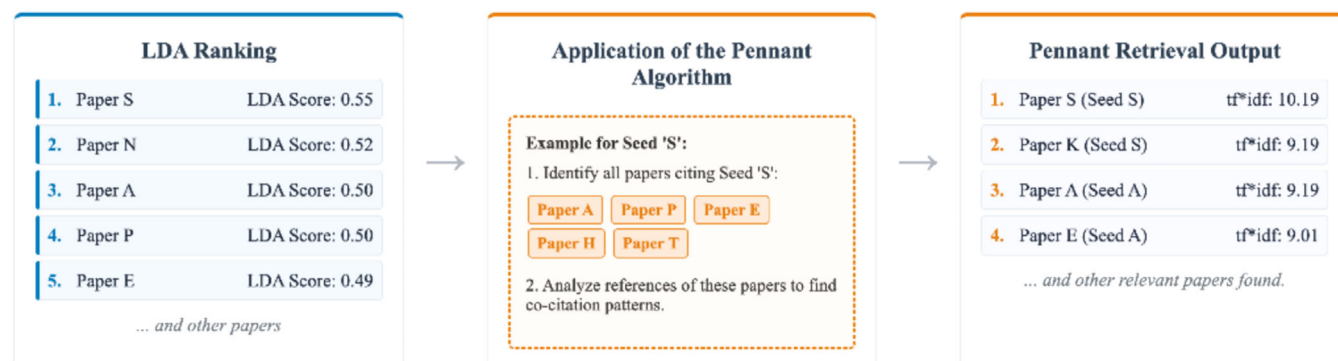


FIGURE 3 Application of the pennant retrieval algorithm (for interactive version, see https://mugeakbulut.com/incremental_refinement/fig3.html).

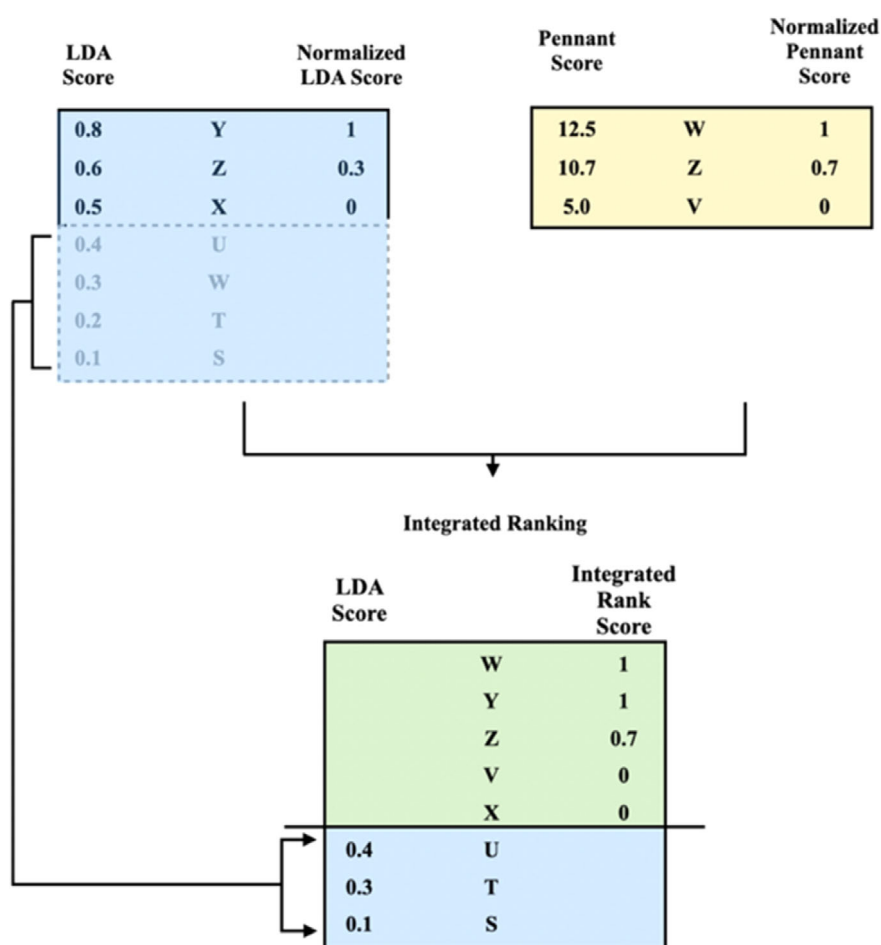


FIGURE 4 Ranking integration.

3.2.4 | Integration of rankings

The LDA algorithm produces a similarity score between the query and all papers with summaries, yielding a fixed-length ranking. In contrast, the pennant-based list varies in length depending on co-citation frequency. To integrate them, the pennant ranking is used as the base, and both

lists are equalized in length through interpolation or truncation. Scores from each are then normalized using min-max normalization and merged (Figure 4).

We used Formula 3 for normalization:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

The min-max values are converted to 0 and 1, respectively, and all others are transformed into decimal numbers in between (Thara et al., 2019). Papers appearing in both rankings are assigned the higher of their two normalized scores in the integrated list (e.g., Paper Z). For papers lacking summaries (e.g., Paper V), no LDA scores can be calculated. Yet, they appear in the integrated list via pennant retrieval if co-citation data exists. Once the list reaches its full length, remaining high-scoring papers from the original LDA list that were initially excluded are appended to the end, beginning with the highest ranked (e.g., Paper U onward in Figure 4). This approach ensures topical relevance is preserved while filling the tail end of the list with strong LDA-based candidates.

3.2.5 | Personalization of relevance rankings

The integrated ranking is adjusted to prioritize either highly relevant or diverse papers depending on user preferences. Users can modify the weights to emphasize either the pennant retrieval or LDA scores (or some combination thereof). When papers appear in both lists, their scores in the personalized version reflect the emphasized algorithm rather than the maximum of both normalized scores.

3.3 | Platform interface and user interaction

A key contribution of this study is LuminaCite, a prototype platform operationalizing the incremental refinement model through an interactive interface. It provides a dynamic search environment where users can control the balance between relevance and diversity using a slider and apply various filters (Figure 5). Search results are visualized through interactive pennant diagrams and topic distribution graphs, enabling users to explore the semantic and citation structure of the literature. When a subject is selected from the pennant diagram, relevant papers are highlighted with bold borders and dark colors.

The system defaults to $w = 0.5$ (e.g., a 50–50 relevance–diversity ratio) unless otherwise specified. The ranking is updated in real time as users adjust the slider. Each result includes metadata, LDA, pennant, and integrated scores (Figure 6). LuminaCite provides both an intuitive and controlled IR experience, enabling users to expand their queries and discover new connections.

3.4 | Performance evaluation

Both relevance and diversity were evaluated. As only a small fraction of papers in the iSearch dataset (<2%) include relevance assessments, degree centrality within the citation network was used as a proxy for relevance (Giustolisi et al., 2020; Marujo et al., 2017; Ribeiro & de Matos, 2011). Centrality scores were calculated within each paper's primary arXiv subject category using the NetworkX Python package. To assess ranking performance, Discounted Cumulative Gain (DCG) and normalized DCG (NDCG) metrics were applied.

DCG measures the performance of a ranking by summing the relevance scores of papers, with each score discounted by the logarithm of its rank, assuming that highly relevant papers are more useful when they appear earlier in the ranking and thus given more weight. The DCG at a particular rank position p is formally defined as:

$$DCG = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} \quad (4)$$

where rel_i is the relevance score of the paper at position i .

To allow for the comparison of rankings of different lengths, DCG is normalized (NDCG) by dividing the DCG score by the Ideal DCG (IDCG), which is the DCG score of the best possible ranking up to position p (Formula 5). An NDCG score of 1.0 indicates a perfect ranking (Schuler et al., 2022).

$$NDCG_p = DCG_p / IDCG_p \quad (5)$$

NDCG is one of the most stable and sensitive IR performance metrics (Sakai, 2007a). To assess the effectiveness of the integrated method on LDA and pennant retrieval, we applied a bootstrap resampling procedure and drew 10,000 random samples (with replacement) from the query set. For each sample, we calculated the mean NDCG values at various cut-off points along with confidence intervals.

Diversity was measured by counting the number of unique primary arXiv subject categories present in the top 50 results. As arXiv categories are independent of LDA topics, they help avoid bias in measuring diversity. To benchmark diversity performance, the MMR algorithm was applied post-hoc to the ranking lists produced by LDA, pennant retrieval, and their integration (rather than MMR algorithm functioning as a primary ranking method). This allowed for comparison between a standard method (Shannon diversity index) and the inherent diversity generated by LDA and pennant retrieval algorithms.

big bang

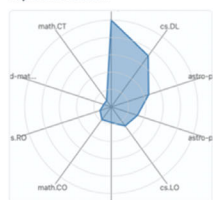


Subject Categories

Search subject...

- ☐ Astrophysics (astro-ph) (105378)
- ☐ High Energy Physics - Phenomenology (hep-ph) (76298)
- ☐ High Energy Physics - Theory (hep-th) (68867)
- ☐ General Relativity and Cosmology (gr-qc) (32746)

Topic Distribution



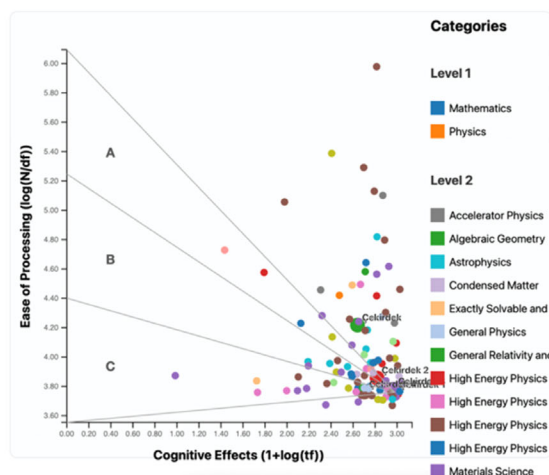
- ☒ astro-ph (22)
- ☒ cs.DL (16)
- ☒ astro-ph.SR (10)
- ☒ astro-ph.IM (7)
- ☒ cs.LG (6)
- ☒ cond-mat.mtrl-sci (4)
- ☒ math.CO (4)

Apply Filters

Personalize Search Results

Relevance Diversity

By increasing the diversity level, you can access interdisciplinary studies that offer different perspectives, even if they're not directly related to your search topic.



How to Read Pennant Diagrams?

Pennant Diagrams visualize paper relationships along cognitive effects (topical relevance) and ease of processing (access specificity) axes.

- Cognitive Effects (X Axis): Papers to the right are more topically similar to the core paper (high co-citation).
- Ease of Processing (Y Axis): Papers above have more specific connections to the core (high total citation).

Sectors:

- Sector A (Successors): Direct continuations of the core paper - specific and current studies.
- Sector B (Peers): Papers published around the same time as the core, more general or derivative studies.
- Sector C (Predecessors): Papers that form the foundation of the core paper - older and more general studies.

Click on nodes or topics in the diagram to access paper details and topic statistics.

FIGURE 5 Screenshot of the interactive pennant diagram.

An Asymmetric e+e- Collider at the psi'

Daniel M. Kaplan, Bruce C. Brown, Harry N. Nelson

A highly-asymmetric "psi'" factory" may be the best approach for studying D0 anti-D0 mixing.

ID: PN134787

High Energy Physics - Experiment (hep-ex)

High Energy Physics - Phenomenology (hep-ph)

Accelerator Physics (physics.acc-ph)

Integrated: 0.881

LDA: 0.881

Pennnat: 0.501

Measurement of the Masses and Lifetimes of B Hadrons at the Tevatron

P. Catastini (for the CDF Collaboration, D0 Collaboration)

The latest results for the B Hadron sector at the Tevatron Collider are summarized. The properties of B Hadrons can be precisely measured at the Tevatron. In particular we will focus on the masses and lifetimes. The new Tevatron results for the CP violation in B Hadrons will be also discussed.

ID: PN138457

High Energy Physics - Experiment (hep-ex)

Integrated: 0.872

LDA: 0.872

Pennnat: 0.414

Previous 3 4 **5** 6 7 Next

FIGURE 6 Screenshot of the relevance ranking.

MMR algorithm selects papers with both high relevance to the query (Q) and minimally redundant (i.e., diverse), and improves IR performance (Yang et al., 2007). It starts with the paper most relevant to the query and then iteratively adds new ones (D_i) by balancing relevance to the

query (Sim_1) and dissimilarity to previously selected papers (Sim_2). This balance is controlled by the parameter λ ($0 \leq \lambda \leq 1$), where higher values prioritize relevance and lower ones favor diversity. The selection maximizes the following function (Carbonell & Goldstein, 1998, p. 335):

TABLE 2 Comparison of maximal marginal relevance (MMR) and the integration function.

Feature	Integration algorithm	MMR algorithm
Purpose	Incremental refinement via high relevance values from LDA and pennant	Incremental refinement via marginal relevance
Method	Score merging	Systematic re-ranking
Input	Term frequencies (LDA), citations (pennant)	Paper-to-paper similarity and query-to-paper similarity
Personalization	Weighting of pennant and LDA algorithms	λ (lambda) control

Abbreviations: LDA, latent dirichlet allocation; MMR, maximal marginal relevance.

$$MMR = \text{Arg max} \{ D_i \in R|S \} \left[\lambda \times \text{Sim}_1(D_i, Q) - (1 - \lambda) \times \max \{ D_j \in S \} \text{Sim}_2(D_i, D_j) \right] \quad (6)$$

where $\text{Sim}_1(D_i, Q)$ is the relevance of paper D_i to query Q ; $\text{Sim}_2(D_i, D_j)$ is the similarity between D_i and already selected paper D_j ; $R|S$ is the set of remaining (unselected) papers; and λ adjusts the relevance–diversity trade-off.

MMR's goal of creating a relevant and diverse result set closely aligns with that of the integration algorithm. However, their approaches differ significantly: MMR reduces redundancy through an iterative penalization process based on a greedy approximation, which implicitly assumes that similar papers will cover similar aspects (Santos et al., 2010). In contrast, the proposed integration algorithm promotes diversity more naturally by combining distinct relevance signals (LDA and citation scores) without penalization. Personalization is achieved via the λ parameter in MMR and algorithm weighting in the integration approach (Table 2).

Finally, robustness tests were conducted to assess whether performance remained stable under alternative conditions. They examined sensitivity to the LDA topic-number setting, tested whether the pennant post-processing step remained effective when LDA was replaced with two neural models, and evaluated robustness to query reformulation using paraphrased variants of the original iSearch queries. Together, these analyses extend the performance evaluation beyond the main ranking metrics and provide a broader basis for interpreting the stability of the proposed framework.

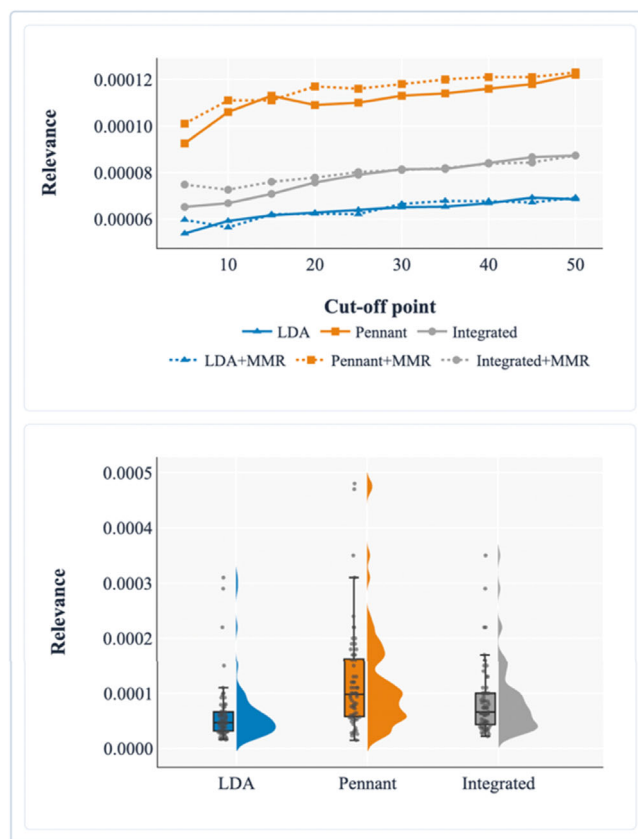


FIGURE 7 Comparison of algorithms based on relevance values.

4 | FINDINGS AND DISCUSSION

The optimal relevance ranking developed in this study is designed to include both highly relevant and thematically diverse papers. To achieve this, LDA and pennant retrieval algorithms were combined to generate rankings prioritizing relevance and capturing topical diversity. Rather than assigning a single overall score to each ranking, we analyzed their specific characteristics.

4.1 | Features emphasized by algorithms

The first 50 results of LDA, pennant, and incrementally refined integrated retrieval algorithms were evaluated for relevance and subject diversity. Figure 7 shows average relevance values across 65 queries at various cut-off points. Relevance is calculated from the degree centrality (centrality-as-relevance) scores of papers in the iSearch collection, summing to 1 (Ribeiro & de Matos, 2011, p. 280). Figure 8 presents Shannon diversity index values.

Pennant retrieved more relevant papers across all cut-off points, while LDA ranked more topically diverse papers higher. The integrated ranking captured the

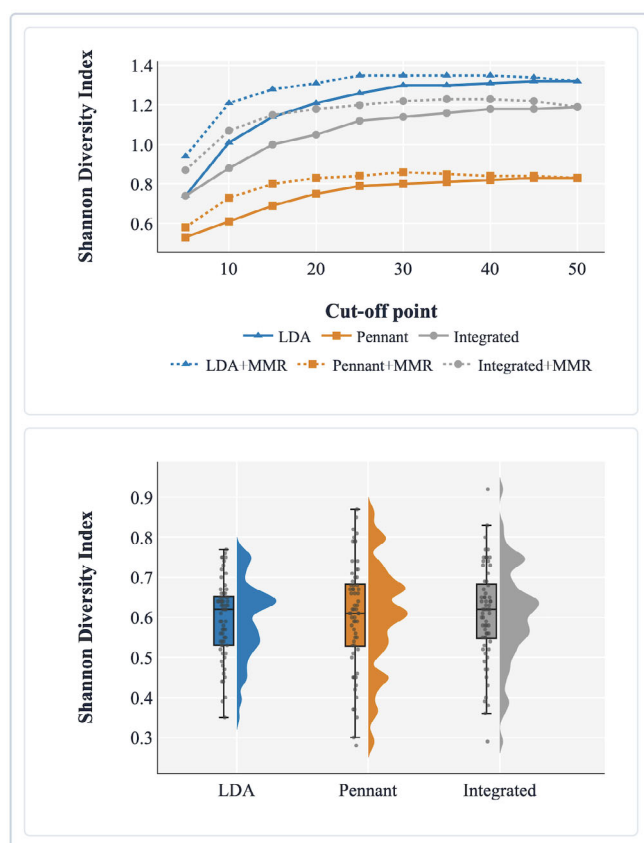


FIGURE 8 Comparison of algorithms based on Shannon diversity index. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig7-8.html.

strengths of both: retrieving highly relevant papers and expanding topical diversity.

This combined ranking mirrors MMR's conceptual approach of balancing relevance and diversity. For validation, MMR was applied to all three rankings with the relevance-diversity balance ($\lambda = 0.5$). Dotted lines in Figures 7 and 8 reflect the outcomes after MMR re-ranking.

MMR applied to pennant retrieval improved relevance at all cut-off points but 15 by pushing similar papers lower and elevating topically distinct yet still relevant ones. As the pennant algorithm tends to load papers on similar topics first, relevance scores generally decline more steeply as cut-off thresholds increase. MMR counteracts this by promoting marginal yet relevant papers to the top. The small increase in relevance at cut-off point 15 for pennant retrieval (Figure 7) reflects this dynamic: highly co-cited relevant papers are elevated in the ranking even if they are on the same topic. They can still rise in the rankings due to their strong co-citation scores, as this is not penalized in pennant retrieval, unlike MMR.

When applied to the integrated ranking, MMR improved relevance in the top 25 results. Beyond that, the difference diminished, likely due to the influence of

pennant retrieval in the integrated model placing relevant papers on similar topics at the top before adding marginal ones. Since the iSearch corpus has low citation density (0.0021), marginal papers appear earlier. In denser corpora, such a shift would likely occur at later cut-off points.

Regarding diversity (Figure 8), LDA outperformed the others due to its word-based similarity model. Pennant retrieved highly relevant papers with strong topical cohesion. As unrelated papers rarely co-cite the same papers, diversity in pennant-based rankings is lower, unless cited papers span multiple disciplines (Waltman & Van Eck, 2012).

However, applying MMR increased diversity across all methods and cut-off points. Pennant retrieval, despite its initial topical tightness, brought boundary-spanning papers with cross-disciplinary connections into top ranks when MMR promoted them based on high relevance but low redundancy. These often-overlooked marginal papers tend to be the most impactful for cross-disciplinary discovery.

In summary, LDA supports topic diversity, while pennant retrieval emphasizes relevance through citation patterns. Their integration produces rankings that balance core and marginal relevance, enhancing both topical breadth and discovery potential.

4.2 | Overall performance evaluation of rankings: DCG and NDCG

IR metrics should reward algorithms retrieving most relevant papers early in the ranked list (Giner, 2022; Sakai, 2007b). To assess ranking performance, we used DCG and its normalized form (NDCG) that accounts for the ideal ranking to make comparisons across different algorithms. DCG evaluates the concentration of relevant papers at the top of a ranked list. DCG values are divided by the ideal relevance ranking to obtain normalized values within the 0–1 range based on centrality-as-relevance values of papers.

The pennant retrieval algorithm yields the highest average DCG scores across 65 queries, indicating its strength in retrieving highly relevant papers in the top ranks (Figure 9). However, the integrated ranking outperforms all others when measured by NDCG in terms of relative closeness of each ranking to the ideal one. Higher NDCG values suggest that the proposed integrated algorithm offers a ranking structure that more closely mirrors the ideal relevance distribution.

Detailed analysis at various cut-off points (Figure 10) further illustrates this trend. DCG values increase across all algorithms as more papers are

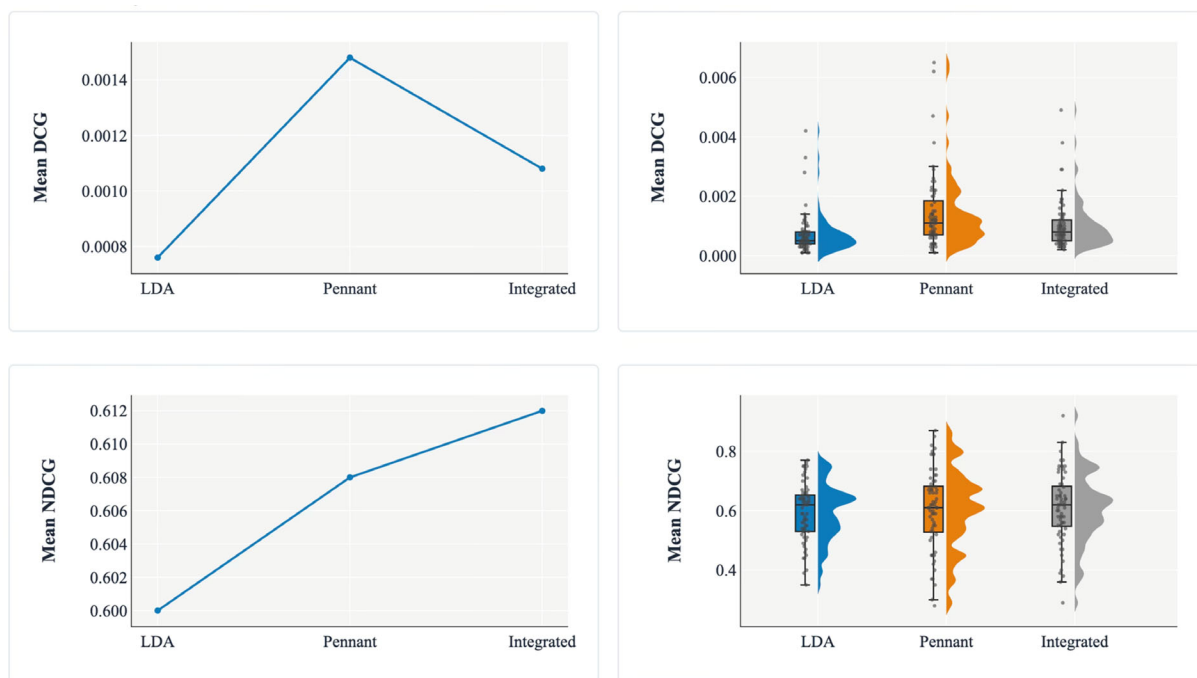


FIGURE 9 Average discounted cumulative gain (DCG) (top) and normalized discounted cumulative gain (NDCG) (bottom) values of algorithms. Interactive version available at: https://mugeakbulut.com/data/incremental_refinement/fig9.html.

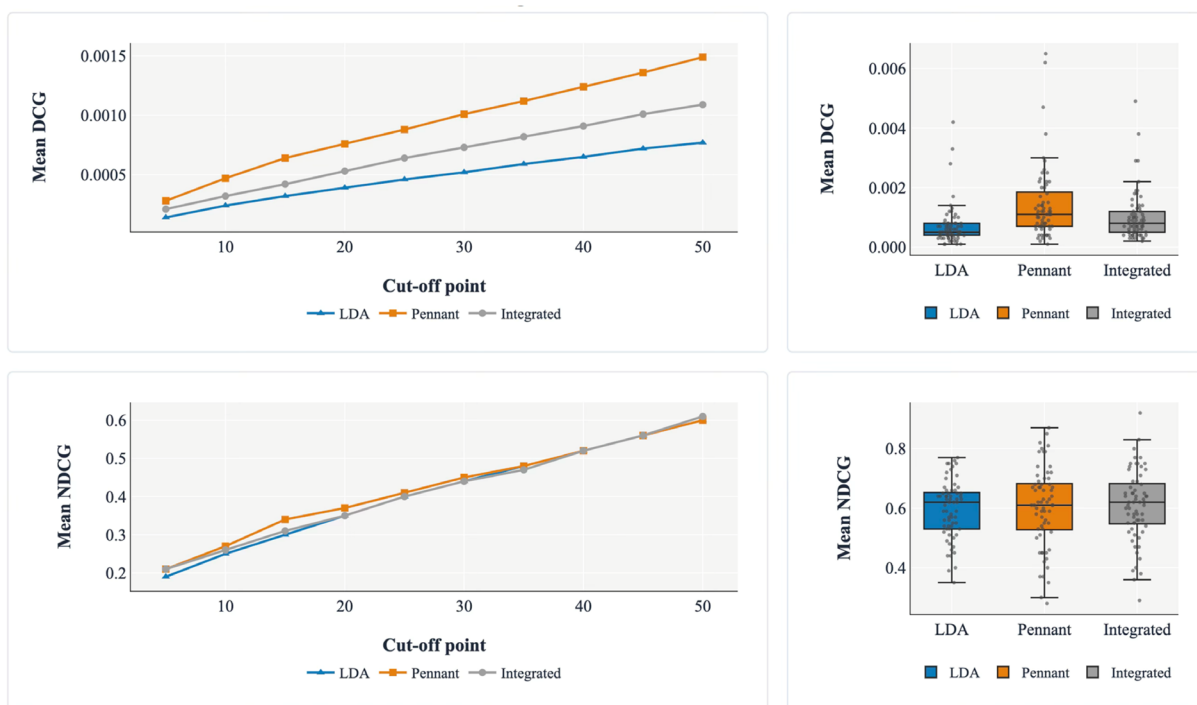


FIGURE 10 Average discounted cumulative gain (DCG) (top) and normalized discounted cumulative gain (NDCG) (bottom) values of algorithms at various cut-off points. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig10.html.

retrieved (top), but the integrated model consistently offers relevant papers at higher ranks. Its advantage is likely due to combining the strengths of both methods: like pennant retrieval, it places highly relevant papers

at the top, and, with support from LDA, it retains moderately relevant yet topically diverse papers at reasonable ranks. This yields a more balanced relevance curve closer to the ideal.

The limitations of the standalone algorithms are also evident: pennant retrieval tends to load highly relevant but thematically homogeneous papers in the top ranks, while LDA seems to sacrifice relevance for diversity. Both tendencies depress NDCG values. The integrated model improves normalized gain by balancing the two and delivers a more relevant and more stable ranking closer to the ideal one. However, a bootstrap analysis comparing the mean NDCG values of three algorithms did not produce statistically significant differences (https://mugeakbulut.com/data/LDA_pennant/paired_bootstrap_10k.html). This outcome is likely attributable to inherent dependencies across ranked lists and queries. The ranked lists for each query share topic-model cues and citation-based features. The pairs of performance values for each query are not statistically independent (Wilbur, 1994, p. 271). These violate the independence assumptions that resampling-based inference relies on. Moreover, the distributions of NDCG values across retrieval methods are not identical (Figures 9 and 10).

Within such contexts, bootstrap resampling may not yield meaningful inferential distinctions beyond what descriptive comparisons already convey. Given this, we focus on descriptive analysis of performance trends such as topical overlap, coverage, novelty, the effects of co-citation density, and algorithmic behaviors of retrieval methods.

4.3 | Topical overlap, coverage and novelty

We also evaluated performance using “coverage” and “novelty” metrics (Figure 11), which measure their success in accessing both known relevant and previously

unknown but potentially relevant topics, respectively. We considered the topics of papers cited in the bibliographies of the seed papers as “known relevant”, otherwise “unknown but potentially relevant” in the first 50 results.

The integrated ranking and LDA show similar median coverage values, both outperforming pennant retrieval. The integrated model exhibits a bimodal distribution concentrated around 75% and 100% coverage, suggesting that combining LDA and pennant retrieval produces rankings with known relevant topics more effectively and covers query topics more broadly.

For novelty, they achieved comparable median values (50%). Occasional dips to 0% indicate that some rankings are highly focused on core topics, especially in their top results. That the pennant retrieval algorithm occasionally reaches a 100% novelty ratio highlights its strength to uncover marginal topics through sparse co-citation networks. LDA and the integrated model tend to hover around 90%, indicating a more consistent but slightly narrower range of topical discovery.

4.4 | Effect of co-citation density and algorithm behavior

Co-citation density plays a critical role in shaping the behavior of the pennant retrieval algorithm and, by extension, the integrated ranking. To explore this, we examined the effect of co-citation count (tf) by grouping queries into two categories: $tf \leq 5$ and $tf > 5$. Co-citation density has opposing effects on relevance and diversity (Figures 12 and 13). The integrated ranking achieves higher relevance but lower diversity in relatively higher co-citation density environments ($tf > 5$). Conversely, diversity improves while relevance diminishes in sparser

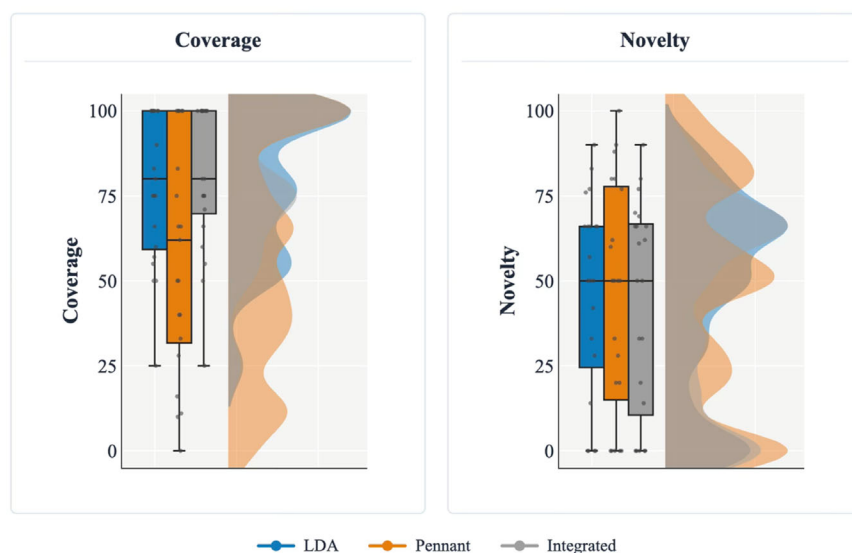


FIGURE 11 Coverage and novelty values. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig11.html.

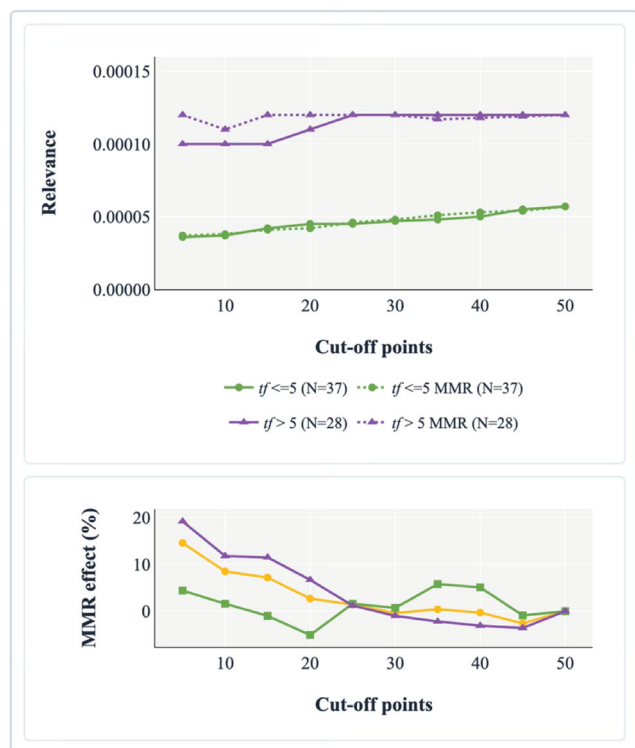


FIGURE 12 Relevance values and maximal marginal relevance (MMR) effect for the integrated ranking (tf threshold ≤ 5).

ones ($tf \leq 5$). The effect of the MMR algorithm also varies with co-citation density, boosting relevance primarily in the top ranks in high-density contexts, but influencing a broader portion of the ranking in low-density cases.

Contributions of LDA and pennant retrieval to the integrated ranking shift modestly with co-citation density (Figure 14). In lower-density environments ($tf \leq 5$), both algorithms contribute similarly (median = 25). In higher-density environments ($tf > 5$), pennant retrieval contributes slightly more (median = 27.5) than LDA (22.5), indicating that pennant retrieval exhibits a more pronounced overall influence across the first half of the list where co-citation patterns are more informative. Conversely, in sparser co-citation settings, LDA maintains greater influence by establishing broader thematic breadth. The integrated model demonstrates adaptability to varying citation conditions, preserving the complementary strengths of both approaches. Notably, even in corpora with low citation density such as iSearch, pennant retrieval still contributes effectively by identifying both core and marginally relevant papers that LDA may overlook.

4.5 | Sensitivity and robustness tests

We examined the proposed model's sensitivity to LDA hyperparameters and its robustness to alternative first-stage

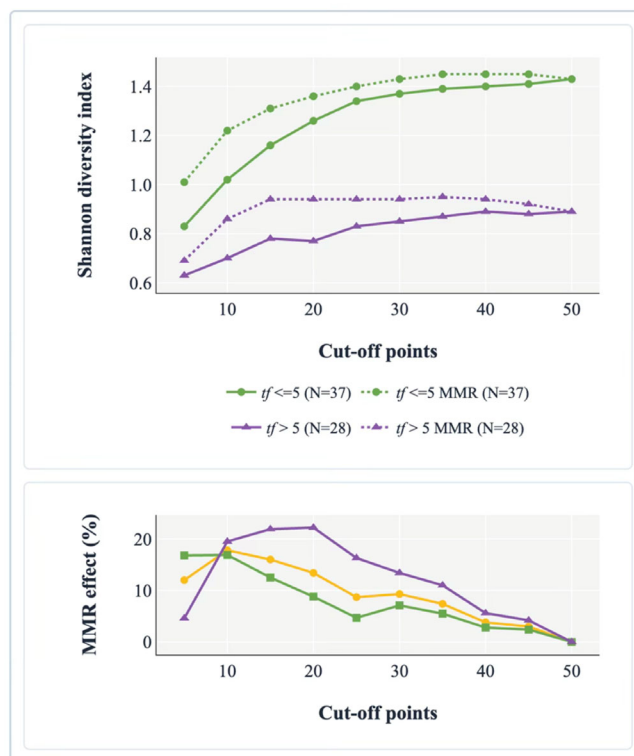


FIGURE 13 Diversity values and maximal marginal relevance (MMR) effect for integrated ranking (tf threshold > 5). For interactive version, see https://mugeakbulut.com/incremental_refinement/fig12-13.html.

retrievers and query variants. As noted earlier, the number of topics (k) is the most consequential LDA parameter. We identified $k = 130$ as the best-supported configuration and stress-tested whether the post-processing step depends on this choice by also evaluating $k = 50$ and $k = 200$.

As expected, LDA's absolute performance varies with k (Figure 15). The best-supported setting ($k = 130$) yielded the highest median NDCG@10 (0.000103), whereas the coarser configuration ($k = 50$) produced a substantially lower baseline (0.000023), and the finer one ($k = 200$) fell in between (0.000049). The pennant post-processing step produced a stable positive effect across all three settings and cut-off points.

Second, we examined whether the pennant post-processing step generalizes beyond LDA-based seeding by replacing the first-stage retriever with two neural models while keeping the downstream citation-based procedure unchanged. SBERT encodes documents as dense semantic vectors optimized for general textual similarity (Reimers & Gurevych, 2019), whereas SPECTER incorporates citation information during pretraining (Cohan et al., 2020). Both models were applied to the full corpus, and the top five retrieved papers per query were used as seeds in place of LDA-derived ones. This design isolates post-processing from the choice of first-stage

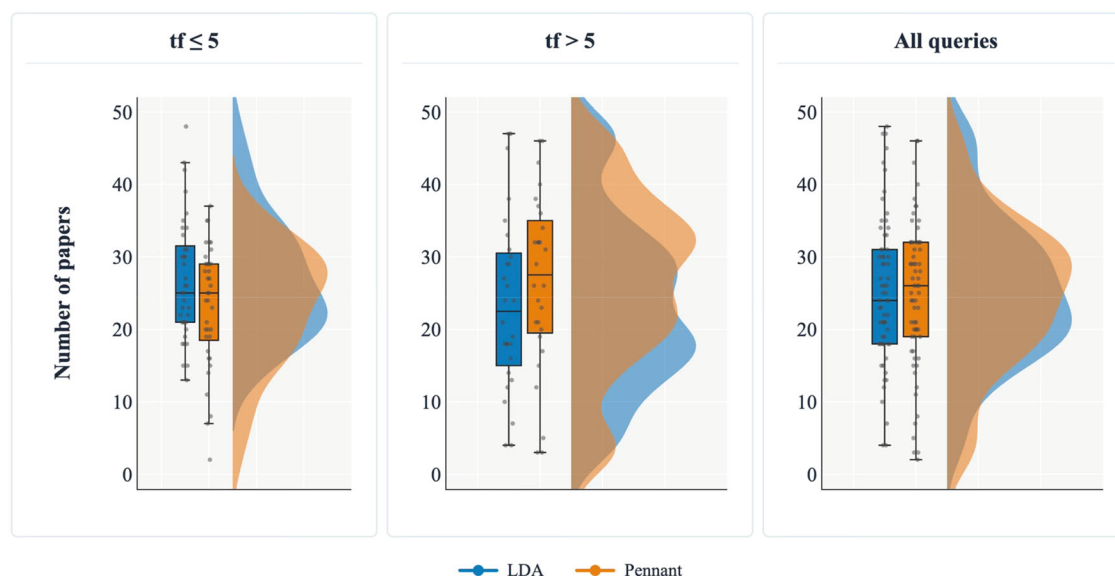


FIGURE 14 Contributions of latent dirichlet allocation (LDA) and pennant retrieval to the integrated ranking across co-citations. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig14.html.

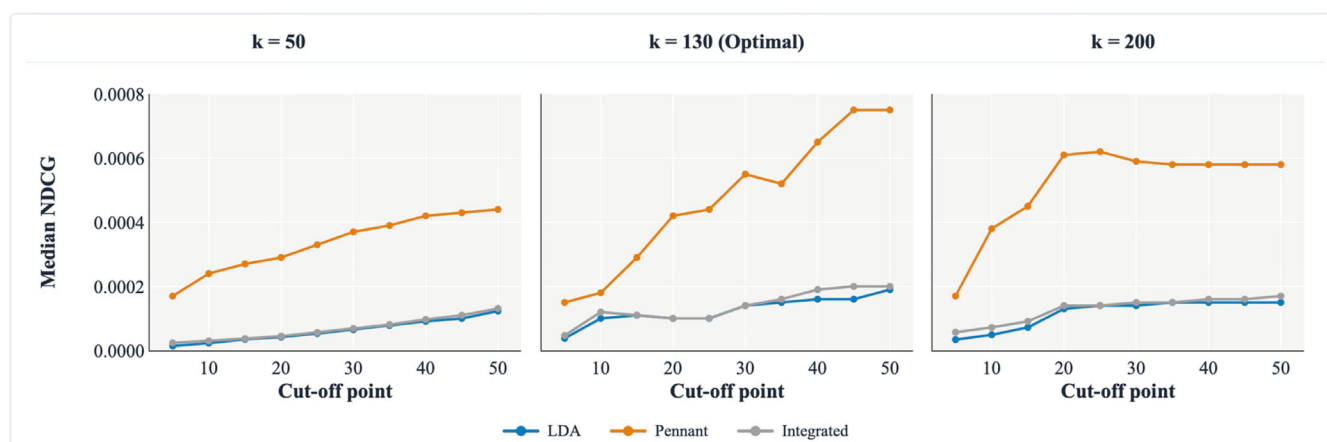


FIGURE 15 Effect of latent dirichlet allocation (LDA) hyperparameter k on retrieval performance. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig15.html.

retriever; any performance differences therefore reflect variation in seed sets rather than changes in the downstream integration logic.

Pennant post-processing consistently improves the mean NDCG across all cut-off points for both neural retrievers (Figure 16), supporting the claim that the pennant contribution is largely model-agnostic at the post-processing stage (Query 10, an outlier, is excluded from the SBERT panel). Moreover, it relies on a transparent $tf \cdot idf$ -style formulation applied to citation counts and can be visualized via pennant diagrams, whereas the similarity scores of SBERT and SPECTER derive from high-dimensional learned representations and are less directly interpretable. Thus, pennant post-processing preserves the explainability advantages of citation-based reasoning

even when paired with dense neural retrievers. Although absolute performance depends on the quality of the initial retriever, the citation-based refinement step remains effective across conceptually different seed generators. Taken together, Figures 15 and 16 indicate robustness to both within-model hyperparameter variation and changes in the first-stage retrieval process.

Third, we assessed sensitivity to query variation, a key dimension of out-of-distribution robustness (Liu et al., 2023, 2025; Penha et al., 2022). We used users' natural language statements for each iSearch topic, rather than the formal search terms used earlier, and generated paraphrased variants using ChatGPT without altering semantic content. Each variant was run on the iSearch corpus and evaluated using the mean NDCG

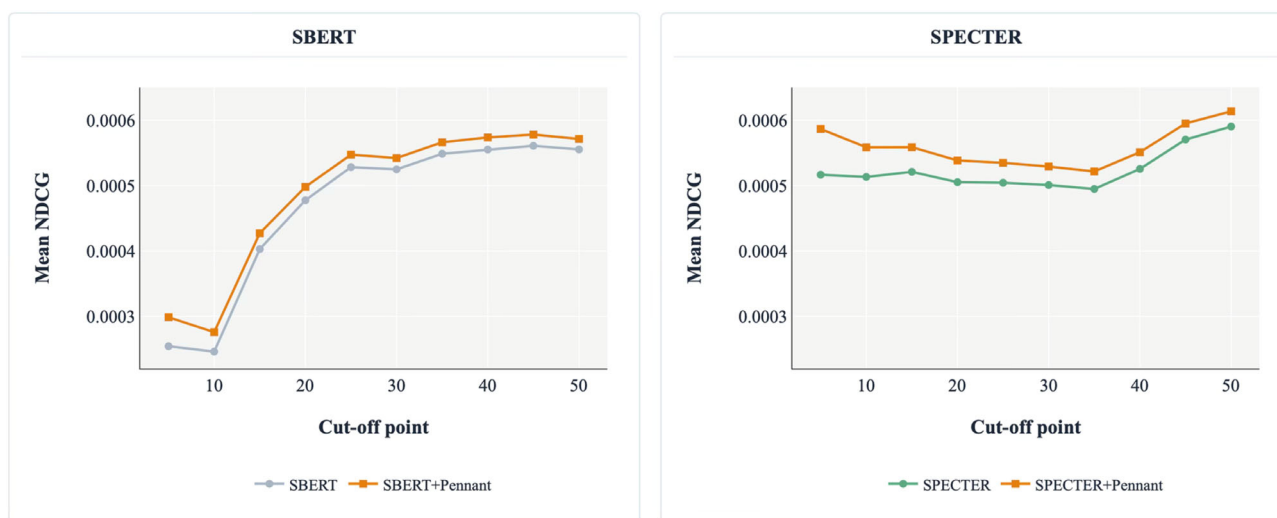


FIGURE 16 Pennant post-processing effect on SBERT and SPECTER. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig16.html.

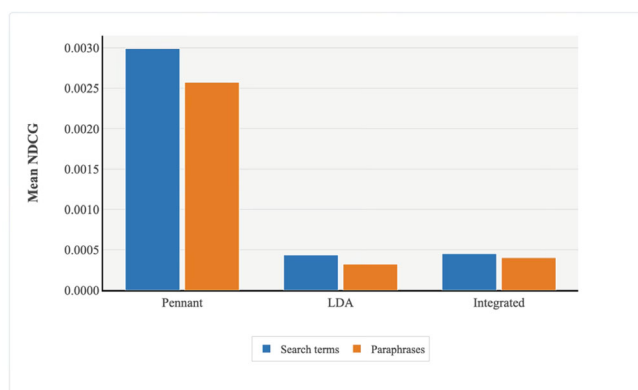


FIGURE 17 Robustness to query reformulation across 65 queries. For interactive version, see https://mugeakbulut.com/incremental_refinement/fig17.html.

performance averaged across cut-off points 5, 10, 20, and 50 for the original and paraphrased query variants. As expected, performance declined for all methods (Figure 17), LDA being more sensitive to query variants (−26.0%) than pennant (−14.0%) and the integrated method (−10.9%). Pennant retrieval achieved substantially higher performance than the other methods, suggesting that citation-based signals contribute to robustness under paraphrasing. However, the final ranking still reflects the behavior of the lexical first-stage retriever.

Overall, the proposed model performs robustly under suboptimal LDA hyperparameters, remains largely agnostic to the choice of first-stage neural retriever, and responds relatively well to query variation.

5 | CONCLUSION AND FURTHER RESEARCH

This study explored the potential of integrating probabilistic topic modeling (LDA) with pennant retrieval, a bibliometric approach based on RT, to create rankings optimized for both relevance and diversity. The pennant algorithm consistently retrieves papers on similar and query-relevant topics in the top ranks, followed by marginal yet potentially cross-disciplinary ones supporting discovery. In contrast, LDA relying on word distributions places marginally relevant papers at the top. Integrating them produces relevance rankings that include both core and boundary-spanning papers, increases relevance compared to LDA alone, and effectively broadens topical coverage.

The integrated ranking method preserves the unique features and contributions of each algorithm. In LDA, words with similar probabilities across topics are less informative. Distinctiveness emerges when probabilities are high for one (or a few) topics and low for others. In pennant retrieval, discriminability is higher for papers with similar total citation and co-citation counts, especially when the latter is relatively high. After normalization and merging, the integrated model prioritizes papers with discriminative characteristics in both.

The resulting ranking reflects the prevailing structures in the data. A strong probabilistic word distribution increases the influence of LDA and plays a greater role in providing thematic diversity in sparse citation environments. In contrast, a corpus with higher co-citation density gives greater weight to pennant retrieval.

The integrated method thus balances word frequency and citation information while preserving each algorithm's distinctive behavior. It adjusts the ranking dynamically based on co-citation density of the corpus.

Overall, the integrated method demonstrated more consistent performance closer to the ideal relevance distribution than either algorithm alone. Unlike MMR, which relies on similarity penalization as a novelty-centric strategy and may be insufficient for consistent diversification on its own (Santos et al., 2012), the proposed integration method remains neutral and preserves the distinctive features of each algorithm without applying penalties. It also covers known relevant topics more thoroughly.

An additional contribution is methodological: Limiting LDA analysis only to titles and abstracts significantly reduces computational time while maintaining effectiveness. Combining summary texts with bibliometric data offers an interpretable, customizable, and discovery-oriented alternative for literature reviews and exploratory research.

A key limitation of this study is its focus on a single corpus. Citation patterns, abstract lengths, topic structures, and other factors vary across fields (Liu & Fang, 2020; Samraj, 2005). To ensure generalizability, broader validation is needed across domains, using varying datasets in size including citation information, full-texts/abstracts, subject classifications, and expert relevance assessments. In addition, we acknowledge that the bag-of-words nature of LDA cannot fully resolve synonymy or polysemy and may yield low topic coherence in small-sample corpora, while pennant retrieval is constrained by low citation density. Our integration does not fundamentally remove these challenges but provides a transparent, user-controllable mechanism that mitigates them in practice.

The integrated approach shows strong potential but requires further evaluation, including its computational efficiency, dynamism, replicability, and scalability (Ballester & Penner, 2022). LDA presents computational challenges in large and dynamic datasets (Mahajan et al., 1999). Our approach mitigates some of these issues, yet it is critical to test it further on large-scale and dynamic citation indexes (e.g., Web of Science), which require licensing, LDA recalculation, and citation updates. Scalability and replicability should be explored using variable sample sizes and collections.

Moreover, identifying algorithmic biases or representation gaps in the integrated ranking is essential, especially regarding its impact on underrepresented research areas or newly emerging terms. Fairness metrics should be used to strengthen the ethical dimension of the method (Xu et al., 2025).

These directions will inform the design of next generation hybrid IR systems that combine both structured and dynamic topical and bibliometric data with probabilistic reasoning and user-controllable interfaces.

ACKNOWLEDGMENTS

We acknowledge support from TÜBİTAK's National Postdoctoral Research Fellowship Program 2218 (Project no. 124C132). We thank the iSearch Team (Peter Ingwersen, Birger Larsen, Haakon Lund, and Marianne Lykke) for the test collection and the anonymous reviewers for constructive comments.

We used ChatGPT to improve the language of the manuscript and take full responsibility for its content.

DATA AVAILABILITY STATEMENT

The iSearch dataset is licensed and not publicly available. Access may be granted upon reasonable request, subject to license terms. Code to download metadata from [arXiv.org](https://arxiv.org/help/api) via its API (<https://arxiv.org/help/api>): https://mugeakbulut.com/phd/codes/iSearch_verilerini_Arxivden_indirme.py. LuminaCite (v.1) prototype platform at GitHub repository (requires Python 3.7+): <https://github.com/mugeakbulut/LuminaCite>. MMR re-ranking script at Google Colab notebook: <https://colab.research.google.com/drive/1dESqgDRL6WfyCSDgHxAa1kPD0HKBfK54>.

ORCID

Müge Akbulut  <https://orcid.org/0000-0003-0026-6485>

Yaşar Tonta  <https://orcid.org/0000-0003-0285-1338>

REFERENCES

- Abramo, G., D'Angelo, C. A., & Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics*, 12(4), 1182–1193. <https://doi.org/10.1016/j.joi.2018.09.001>
- Adomavicius, G., & Kwon, Y. (2011). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5), 896–911. <https://doi.org/10.1109/TKDE.2011.15>
- ADS Team. (2008). SAO/NASA ADS abstract service stopword list. <https://ui.adsabs.harvard.edu>
- Agrawal, A., Wei, F., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98, 74–88. <https://doi.org/10.1016/j.infsof.2018.02.005>
- Akbulut, M. (2016). *Atıf klasiklerinin etkisinin ve ilgililik sıralamalarının pennant diyagramları ile analizi*. [Master's thesis, Hacettepe University]. http://www.mugeakbulut.com/yayinlar/Muge_Akbulut_YL_Tez.pdf
- Akbulut, M. (2022). *Bilgi Erişimde İlgili Sıralamalarının Artırımı Olarak Geliştirilmesi*. [Doctoral dissertation, Hacettepe University]. http://www.mugeakbulut.com/yayinlar/Muge_Akbulut_PhD_Tez.pdf

- Akbulut, M., & Tonta, Y. (2022). İlgi sıralamalarının artırımı olarak geliştirilmesi: Pennant erişimle desteklenen yeni bir yöntem önerisi. *Türk Kutuphaneciliği*, 36(2), 169–203. <https://doi.org/10.24146/tk.1062751>
- Akbulut, M., Tonta, Y., & White, H. D. (2020). Related records retrieval and pennant retrieval: An exploratory case study. *Scientometrics*, 122(2), 1103–1133. <https://doi.org/10.1007/s11192-019-03303-9>
- Alaofi, M., Thomas, P., Scholer, F., & Sanderson, M. (2024). LLMs can be fooled into labelling a document as relevant: Best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 annual international ACM SIGIR conference on research and development in information retrieval in the Asia Pacific region, SIGIR-AP 2024* (pp. 32–41). ACM. <https://doi.org/10.1145/3673791.3698431>
- Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press. <https://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf>
- Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, 16(1), 101224. <https://doi.org/10.1016/j.joi.2021.101224>
- Bascur, J. P., Verberne, S., van Eck, N. J., & Waltman, L. (2023). Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. *Scientometrics*, 128(5), 2895–2921. <https://doi.org/10.1007/s11192-023-04681-x>
- Bascur, J. P., Verberne, S., van Eck, N. J., & Waltman, L. (2025). Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks. *Scientometrics*, 130(2025), 1181–1199. <https://doi.org/10.1007/s11192-024-05218-6>
- Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: An introductory overview. In B. Larsen & J. Leta (Eds.), *Proceedings of the 12th international conference on Scientometrics and Informetrics* (Vol. 1, pp. 230–241). International Society for Scientometrics and Informetrics. https://www.issi-society.org/proceedings/issi_2009/ISSI2009-proc-vol1_Aug2009_batch2-paper-1.pdf
- Beel, J., Gipp, B., Langer, S., & Breitingner, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. <https://doi.org/10.1007/s00799-015-0156-0>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Bogers, T., Gäde, M., Hall, M., Koolen, M., Petras, V., & Skov, M. (2025). Exploring the zero-shot known-item retrieval capabilities of LLMs for casual leisure information needs. In *CHIIR '25: ACM SIGIR conference on human information interaction and retrieval* (pp. 316–323). ACM. <https://doi.org/10.48550/arXiv.1002.4665>
- Boyd-Graber, J., & Blei, D. M. (2010). *Syntactic topic models*. arXiv. <https://doi.org/10.48550/arXiv.1002.4665>
- Bradley, K., & Smyth, B. (2001). Improving recommendation diversity. In D. O'Donoghue (Ed.), *Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science* (pp. 141–152). NUIM Department of Computer Science. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.5232&rep=rep1&type=pdf>
- Breuer, T., Frihat, S., Fuhr, N., Lewandowski, D., Schaer, P., & Schenkel, R. (2025). Large language models for information retrieval: Challenges and chances. *Datenbank Spektrum*, 25, 1–11. <https://doi.org/10.1007/s13222-025-00503-x>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418fbfb8ac142f64a-Paper.pdf
- Caballero, K. L., & Akella, R. (2012). Incorporating statistical topic information in relevance feedback. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (p. 1093). ACM. <https://doi.org/10.1145/2348283.2348485>
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48–57. <https://doi.org/10.1109/MCI.2014.2307227>
- Cao, T., Xia, J., Li, Y., Zhang, S., & Tang, A. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7), 1775–1781. <https://doi.org/10.1016/j.neucom.2008.06.011>
- Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 335–336). ACM. <https://doi.org/10.1145/290941.291025>
- Carevic, Z., & Mayr, P. (2014). *Recommender systems using pennant diagrams in digital libraries*. arXiv. <https://doi.org/10.48550/arXiv.1407.7276>
- Carevic, Z., & Schaer, P. (2014). On the connection between citation-based and topical relevance ranking: Results of a pre-test using iSearch. In *Proceedings of the first workshop on bibliometric-enhanced information retrieval* (pp. 37–44). CEUR-WS. <https://ceur-ws.org/Vol-1143/paper5.pdf>
- Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1), 1–50. <https://doi.org/10.1145/2071389.2071390>
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 22 (pp. 288–296). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf>
- Chen, M., & Décary, M. (2018). A cognitive-based semantic approach to deep content analysis in search engines. In *2018 IEEE 12th international conference on semantic computing (ICSC)* (pp. 131–139). IEEE. <https://doi.org/10.1109/ICSC.2018.00027>

- Chen, Z., & Liu, B. (2014). Mining topics in documents: Standing on the shoulders of big data. In S. Macskassy, C. Perlich, J. Leskovec, W. Wang, & R. Ghani (Eds.), *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1116–1125). ACM. <https://doi.org/10.1145/2623330.2623622>
- Chimah, J. N., & Ude, F. I. (2020). Current trends in information retrieval systems: Review of fuzzy set theory and fuzzy Boolean retrieval models. *Journal of Library Services and Technologies*, 2(2), 48–56. <https://doi.org/10.47524/jlst.v2i2.5>
- Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. *arXiv*. <https://doi.org/10.48550/arXiv.2004.07180>
- Cooper, W. S. (1988). Getting beyond Boole. *Information Processing & Management*, 24(3), 243–248. [https://doi.org/10.1016/0306-4573\(88\)90091-X](https://doi.org/10.1016/0306-4573(88)90091-X)
- Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 36(4), 533–550. [https://doi.org/10.1016/S0306-4573\(99\)00072-2](https://doi.org/10.1016/S0306-4573(99)00072-2)
- Crestani, F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Information Retrieval*, 4(2), 131–152. <https://doi.org/10.1023/A:1009973415168>
- Croft, W. B. (2002). Combining approaches to information retrieval. In K. Sparck Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 485–498). Morgan Kaufmann. https://doi.org/10.1007/0-306-47019-5_1
- Cunningham, E., Smyth, B., & Greene, D. (2025). Facilitating interdisciplinary knowledge transfer with research paper recommender systems. *Quantitative Science Studies*, 6, 854–875. <https://doi.org/10.1162/qss.a.9>
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84. <https://doi.org/10.3166/dn.17.1.61-84>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Ekinci, E., & İlhan Omurca, S. (2020). Concept-LDA: Incorporating Babelify into LDA for aspect extraction. *Journal of Information Science*, 46(3), 406–418. <https://doi.org/10.1177/0165551519845854>
- Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., & Wachsmuth, H. (2024). Who determines what is relevant? Humans or AI? Why not both? A spectrum of human-AI collaboration in assessing relevance. *Communications of the ACM*, 67(4), 31–34. <https://doi.org/10.1145/3624730>
- Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., Osipov, M., Kholodov, M., Ismagilov, R., & Mohan, S. (2018). Best match: New relevance search for PubMed. *PLoS Biology*, 16(8), e2005343. <https://doi.org/10.1371/journal.pbio.2005343>
- Galbreath, B. L., England, E., Johnson, C. M., & Saulnier Lange, J. (2025). AI-infused discovery environments: Information retrieval boon or overpromised hype? *Information Technology and Libraries*, 44(4), 1–17. <https://doi.org/10.5860/ital.v44i4.17465>
- Ganguly, D., & Jones, G. J. (2018). A non-parametric topical relevance model. *Information Retrieval Journal*, 21(5), 449–479. <https://doi.org/10.1007/s10791-018-9329-y>
- Giner, F. (2022). On the effect of ranking axioms on IR evaluation metrics. In *proceedings of the 2022 ACM SIGIR international conference on the theory of information retrieval (ICTIR'22)* (p. 11). ACM. <https://doi.org/10.1145/3539813.3545153>
- Giustolisi, O., Ridolfi, L., & Simone, A. (2020). Embedding the intrinsic relevance of vertices in network analysis: The case of centrality metrics. *Scientific Reports*, 10, 3297. <https://doi.org/10.1038/s41598-020-60151-x>
- Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—Different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2), 981–998. <https://doi.org/10.1007/s11192-017-2296-z>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of The United States of America*, 101(Suppl. 1), 5228–5235. <https://doi.org/10.1073/pnas.0307752101>
- Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1–30. <https://epub.wu.ac.at/3987/1/topicmodels.pdf>
- Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international conference on information and knowledge management* (pp. 55–64). ACM. <https://doi.org/10.1145/2983323.2983769>
- Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics*, 125(3), 2561–2595. <https://doi.org/10.1007/s11192-020-03721-0>
- Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 602. <https://www.proquest.com/scholarly-journals/psychological-relevance-information-science/docview/216898031/se-2-615>
- Hecking, T., & Leydesdorff, L. (2018). *Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps*. *arXiv*. <https://arxiv.org/pdf/1806.01045.pdf>
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 5–53. <https://doi.org/10.1145/963770.963772>
- Holliger, T. S. (2018). Strategic sourcing via category management: Helping air force installation contracting agency eat one piece of the elephant. [Master's thesis, Air Force Institute of Technology]. <https://apps.dtic.mil/sti/pdfs/AD1056353.pdf>
- Huang, L., Liu, H., He, J., & Du, X. (2016). Finding latest influential research papers through modeling two views of citation links. In F. Li, K. Shim, K. Zheng, & G. Liu (Eds.), *Web technologies and applications. APWeb 2016* (pp. 555–566). Springer. https://doi.org/10.1007/978-3-319-45814-4_45
- Huang, X., Chen, C. A., Peng, C., Wu, X., Fu, L., & Wang, X. (2018). Topic-sensitive influential paper discovery in citation network. In *Advances in knowledge discovery and data mining: 22nd Pacific-Asia conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, proceedings, part II* (Vol. 22, pp. 16–28). Springer International Publishing. <http://www.cs.sjtu.edu.cn/~fu-ly/paper/Topic-sensitive%20Influential%20Paper%20Dis.pdf>

- Jiang, Z., Liu, X., & Gao, L. (2014). Dynamic topic/citation influence modeling for chronological citation recommendation. In *Proceedings of the 7th ACM international conference on web search and data mining* (pp. 15–18). ACM. <https://doi.org/10.1145/2663792.2663795>
- Jin, R., Valizadegan, H., & Li, H. (2008). Ranking refinement and its application to information retrieval. In *Proceedings of the 17th international conference on world wide web* (pp. 397–406). ACM. <https://doi.org/10.1145/1367497.1367552>
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate*. arXiv. <https://doi.org/10.48550/arXiv.2509.04664>
- Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24), 7426–7431. <https://doi.org/10.1073/pnas.1424329112>
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. <https://doi.org/10.1002/asi.5090140103>
- Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N., & Bayer, V. (2017). *Towards effective research recommender systems for repositories*. arXiv. <https://arxiv.org/abs/1705.00578>
- Küçükünç, O., & Ferhatosmanoğlu, H. (2011). λ -Diverse nearest neighbors browse for multidimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 25(3), 481–493. <https://doi.org/10.1109/TKDE.2011.251>
- Küçükünç, O., Saule, E., Kaya, K., & Çatalyürek, Ü. V. (2015). Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology*, 5(4), 1–21. <https://doi.org/10.1145/2668106>
- Lei, M., Wang, J., Chen, B., & Li, X. (2001). Improved relevance ranking in WebGather. *Journal of Computer Science and Technology*, 16(5), 410–417. <https://doi.org/10.1007/bf02948958>
- Lewandowski, D. (2009). Ranking library materials. *Library Hi Tech*, 27(4), 584–596. <https://doi.org/10.1108/07378830911007682>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., Küttler, H., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>
- Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora ($N < 1000$). *Journal of the Association for Information Science and Technology*, 68(4), 1024–1035. <https://doi.org/10.1002/asi.23740>
- Li, C., Feng, H., & de Rijke, M. (2020). Cascading hybrid bandits: Online learning to rank for relevance and diversity. In *Fourteenth ACM conference on recommender systems* (pp. 33–42). ACM. <https://doi.org/10.1145/3383313.3412245>
- Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on machine learning* (pp. 577–584). ACM. <https://doi.org/10.1145/1143844.1143917>
- Liu, T., & Gu, X. (2023). LDA-BERT based public opinion subject mining analysis of emergencies. In *Second international conference on algorithms, microchips, and network applications (AMNA 2023)*. SPIE. <https://doi.org/10.1117/12.2679263>
- Liu, X. Z., & Fang, H. (2020). A comparison among citation-based journal indicators and their relative changes with time. *Journal of Informetrics*, 14(1), 1–17. <https://doi.org/10.1016/j.joi.2020.101007>
- Liu, Y. A., Zhang, R., Guo, J., Chen, W., & Cheng, X. (2023). *On the robustness of generative retrieval models: An out-of-distribution perspective*. arXiv. <https://doi.org/10.48550/arXiv.2306.12756>
- Liu, Y. A., Zhang, R., Guo, J., & de Rijke, M. (2025). Robust information retrieval. In *Proceedings of the eighteenth ACM international conference on web search and data mining* (pp. 1008–1011). ACM. <https://doi.org/10.1145/3701551.3703476>
- Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In *Advances in information retrieval. ECIR 2010* (pp. 627–630). Springer. https://doi.org/10.1007/978-3-642-12275-0_63
- Ma, Z., Liu, Y., Yang, Z., Yang, J., & Li, K. (2022). A parameter-free approach to lossless summarization of fully dynamic graphs. *Information Sciences*, 589, 376–394. <https://doi.org/10.1016/j.ins.2021.12.116>
- Mahajan, M., Beeferman, D., & Huang, X. D. (1999). Improved topic-dependent language modeling using information retrieval techniques. In *1999 IEEE international conference on acoustics, speech, and signal processing. Proceedings. ICASSP99* (Vol. 1, pp. 541–544). IEEE. <https://doi.org/10.1109/ICASSP.1999.758182>
- Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. MIT Press. <https://tinyurl.com/2we6zkyk>
- Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3), 216–244. <https://doi.org/10.1145/321033.321035>
- Martinho, V. J. P. D. (2022). Bibliographic coupling links: Alternative approaches to carrying out systematic reviews about renewable and sustainable energy. *Environments*, 9(2), 28. <https://doi.org/10.3390/environments9020028>
- Marujo, L., Ribeiro, R., Gershman, A., De Matos, D. M., Neto, J. P., & Carbonell, J. (2017). Event-based summarization using a centrality-as-relevance model. *Knowledge and Information Systems*, 50(3), 945–968. <https://doi.org/10.1007/s10115-016-0966-4>
- Mayr, P., & Mutschke, P. (2013). Bibliometric-enhanced retrieval models for big scholarly information systems. In *2013 IEEE international conference on big data* (pp. 5–8). IEEE. <https://doi.org/10.1109/BigData.2013.6691762>
- McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on human factors in computing systems* (pp. 1097–1101). ACM. <https://doi.org/10.1145/1125451.1125659>
- Meng, W., Yu, C. T., & Liu, K. L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys*, 34(1), 48–89. <https://doi.org/10.1145/505282.505284>
- Metzler, D., Tay, Y., Bahri, D., & Zhuang, W. (2021). Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, 55(1), 1–20. <https://arxiv.org/pdf/2105.02274>
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9), 810–832.

- [https://doi.org/10.1002/\(SICI\)1097-4571\(199709\)48:9<810::AID-ASI6>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U)
- Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, 10(3), 303–320. [https://doi.org/10.1016/S0953-5438\(98\)00012-5](https://doi.org/10.1016/S0953-5438(98)00012-5)
- Nandy, A., Singh, A., Gupta, V., & Singh, V. K. (2024). Bibliographic coupling and conceptual similarity: Are the bibliographically coupled papers also conceptually similar? *Journal of Scientometric Research*, 13(3), 706–714. <https://doi.org/10.5530/jscires.20041115>
- Năstasă, A., Dumitra, T. C., & Grigorescu, A. (2024). Artificial intelligence and sustainable development during the pandemic: An overview of the scientific debates. *Heliyon*, 10(9), e30412. <https://doi.org/10.1016/j.heliyon.2024.e30412>
- Nguyen, T., & Do, P. (2018). CitationLDA++: An extension of LDA for discovering topics in document network. In *Proceedings of the ninth international symposium on information and communication technology* (pp. 31–37). ACM. <https://doi.org/10.1145/3287921.3287930>
- Pao, M. L. (1993). Term and citation retrieval: A field study. *Information Processing & Management*, 29(1), 95–112. [https://doi.org/10.1016/0306-4573\(93\)90026-A](https://doi.org/10.1016/0306-4573(93)90026-A)
- Penha, G., Câmara, A., & Hauff, C. (2022). Evaluating the robustness of retrieval pipelines with query variation generators. In M. Hagen, et al. (Eds.), *Advances in information retrieval* (Lecture Notes in Computer Science) (Vol. 13185, pp. 393–406). Springer. https://doi.org/10.1007/978-3-030-99736-6_27
- Poetz, M. K., & Prügl, R. (2010). Crossing domain-specific boundaries in search of innovation: Exploring the potential of pyramiding. *Journal of Product Innovation Management*, 27(6), 897–914. <https://doi.org/10.1111/j.1540-5885.2010.00759.x>
- Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, 41(7), 1262–1282. <https://doi.org/10.1016/j.respol.2012.03.015>
- Rafols, I., Porter, A. L., & Leydesdorff, L. (2012). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871–1887. <https://doi.org/10.1002/asi.21368>
- Ravuri, B., Mardis, M. A., Moja, O., Ayinde, L., & Emdad, F. B. (2026). Complex questions and quality answers: Comparing ChatGPT and Gemini as research collaborators. *Journal of the Association for Information Science and Technology*, 77, 795–811. <https://doi.org/10.1002/asi.70056>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Ribeiro, R., & de Matos, D. M. (2011). Centrality-as-relevance: Support sets and similarity as geometric proximity. *Journal of Artificial Intelligence Research*, 42, 275–308. <https://doi.org/10.1613/jair.3387>
- Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294–304. <https://doi.org/10.1108/eb026647>
- Sakai, T. (2007a). On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, 43, 531–548. <https://doi.org/10.1016/j.ipm.2006.07.020>
- Sakai, T. (2007b). On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In *The first international workshop on evaluating information access (EVIA) technologies*, Japan. EVIA 2007 Online Proceedings (pp. 32–43). National Institute of Informatics. <https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf>
- Salton, G., Yang, C. S., & Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. <https://doi.org/10.1145/361219.361220>
- Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24(2), 141–156. <https://doi.org/10.1016/j.esp.2002.10.001>
- Santos, R. L. T., Macdonald, C., & Ounis, I. (2012). On the role of novelty for search result diversification. *Information Retrieval*, 15(5), 478–502. <https://doi.org/10.1007/s10791-011-9180-x>
- Santos, R. L. T., Peng, J., Macdonald, C., & Ounis, I. (2010). Explicit search result diversification through sub-queries. In C. Gurrin, et al. (Eds.), *Advances in information retrieval. ECIR 2010. Lecture notes in computer science* (Vol. 5993, pp. 87–99). Springer. https://doi.org/10.1007/978-3-642-12275-0_11
- Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13), 1915–1933. <https://doi.org/10.1002/asi.20682>
- Schuler, J., Falls, Z., Mangione, W., Hudson, M. L., Bruggemann, L., & Samudrala, R. (2022). Evaluating the performance of drug-repurposing technologies. *Drug Discovery Today*, 27(1), 49–64. <https://doi.org/10.1016/j.drudis.2021.08.002>
- Shen, S., Zhu, D., Rousseau, R., Su, X., & Wang, D. (2019). A refined method for computing bibliographic coupling strengths. *Journal of Informetrics*, 13(2), 605–615. <https://doi.org/10.1016/j.joi.2019.01.012>
- Sjögarde, P., & Ahlgren, P. (2024). Seed-based information retrieval in networks of research publications: Evaluation of direct citations, bibliographic coupling, co-citations, and PubMed-related article score. *Journal of the Association for Information Science and Technology*, 75, 1453–1465. <https://doi.org/10.1002/asi.24951>
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Soboroff, I. (2025). Don't use LLMs to make relevance judgments. *Information Retrieval Research*, 1(1), 29–46. <https://irj.org/article/view/19625/25085>
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell. https://monoskop.org/images/e/e6/Sperber_Dan_Wilson_Deirdre_Relevance_Communication_and_Cognition_2nd_edition_1996.pdf
- Staudinger, M., Kusa, W., & Hanbury, A. (2026). *The LLM effect on IR benchmarks: A meta-analysis of effectiveness, baselines, and contamination*. arXiv. <https://arxiv.org/pdf/2604.05766>

- Swanson, D. R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1), 7–18. <https://doi.org/10.1353/pbm.1986.0087>
- Swanson, D. R. (1986b). Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, 56(4), 389–398. <https://doi.org/10.1086/601800>
- Thara, D. K., PremaSudha, B. G., & Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, 128, 544–550. <https://doi.org/10.1016/j.patrec.2019.10.029>
- Tonta, Y. (1992). Analysis of search failures in document retrieval systems: A review. *The Public-Access Computer Systems Review*, 3(1), 4–53. https://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta_analysis_of_search_failures_in_document_retrieval_systems_a_review.pdf
- Urs, S. (2025). IR meets AI: Looking back, looking forward—A conversation with Stephen Robertson, pioneer of information retrieval. In *Information matters*. Association for Information Science & Technology. <https://informationmatters.org/2025/04/ir-meets-ai-looking-back-looking-forward-a-conversation-with-stephen-robertson-pioneer-of-information-retrieval/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://papers.nips.cc/paper/7181-attention-is-all-you-need>
- Vergoulis, T., Chatzopoulos, S., Kanellos, I., Deligiannis, P., Tryfonopoulos, C., & Dalamagas, T. (2019). BIP! Finder: Facilitating scientific literature search by exploiting impact-based ranking. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2937–2940). ACM. <https://doi.org/10.1145/3357384.3357850>
- Verma, M., Yilmaz, E., & Craswell, N. (2016). On obtaining effort-based judgements for information retrieval. In *Proceedings of the 9th ACM international conference on web search and data mining* (pp. 277–286). ACM. <https://doi.org/10.1145/2835776.2835840>
- Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems*, 22, 1973–1981. <https://proceedings.neurips.cc/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf>
- Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2390–2402. <https://arxiv.org/pdf/1203.0532>
- Wang, X., Zhai, C., & Roth, D. (2013). Understanding evolution of research themes: A probabilistic generative model for citations. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1115–1123). ACM. <https://doi.org/10.1145/2487575.2487698>
- White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory. Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, 58(4), 536–559. <https://doi.org/10.1002/asi.20543>
- White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology*, 58(5), 586–603. <https://doi.org/10.1002/asi.20542>
- White, H. D. (2009). Pennants for Strindberg and Persson. In *Celebrating scholarly communication studies: A festschrift for Olle Persson at his 60th birthday* (pp. 71–83). ISSI. <https://lup.lub.lu.se/search/files/5414075/1459003.pdf#page=74>
- White, H. D. (2010). Some new tests of relevance theory in information science. *Scientometrics*, 83(3), 653–667. <https://doi.org/10.1007/s11192-009-0138-3>
- White, H. D. (2015). Co-cited author retrieval and relevance theory: Examples from the humanities. *Scientometrics*, 102(3), 2275–2299. <https://doi.org/10.1007/s11192-014-1483-4>
- White, H. D. (2016). Bag of works retrieval: TFIDF weighting of co-cited works. In *Proceedings of the 3rd workshop on bibliometric-enhanced information retrieval (BIR2016)* (pp. 63–72). CEUR-WS. <https://ceur-ws.org/Vol-1567/paper7.pdf>
- White, H. D. (2018). Bag of works retrieval: TFIDF weighting of co-cited works with a seed. *International Journal on Digital Libraries*, 19(2–3), 139–149. <https://doi.org/10.1007/s00799-017-0217-7>
- White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327–355. [https://doi.org/10.1002/\(SICI\)1097-4571\(19980401\)49:4<327::AID-ASI4>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4)
- Wilbur, W. J. (1994). Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20, 270–284. <https://doi.org/10.1177/016555159402000405>
- Wu, H. C., Luk, R. W., Wong, K. F., & Kwok, K. L. (2007). A retrospective study of a hybrid document-context based retrieval model. *Information Processing & Management*, 43(5), 1308–1331. <https://doi.org/10.1016/j.ipm.2006.10.009>
- Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3), 1–37. <https://doi.org/10.1145/1361684.1361686>
- Xia, H., Li, J., Tang, J., & Moens, M. F. (2012). Plink-LDA: Using link as prior information in topic modeling. In S. Lee, et al. (Eds.), *Database systems for advanced applications* (pp. 213–227). Springer. https://doi.org/10.1007/978-3-642-29038-1_17
- Xie, X., Liang, Y., Li, X., & Tan, W. (2019). CuLDA_CGS: Solving large-scale LDA problems on GPUs. In *Proceedings of the 24th symposium on principles and practice of parallel programming* (pp. 435–436). ACM. <https://doi.org/10.1145/3293883.3301496>
- Xu, C., Deng, Z., Rus, C., Ye, X., Liu, Y., Xu, J., Dou, Z., Wen, J. R., & de Rijke, M. (2025). FairDiverse: A comprehensive toolkit for fairness- and diversity-aware information retrieval. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval (SIGIR'25)* (pp. 3540–3550). ACM. <https://doi.org/10.1145/3726302.3730280>
- Yang, H. T., Ju, J. H., Wong, Y. T., Shmulevich, I., & Chiang, J. H. (2017). Literature-based discovery of new candidates for drug repurposing. *Briefings in Bioinformatics*, 18(3), 488–497. <https://doi.org/10.1093/bib/bbw030>
- Yang, L., Ji, D., & Leong, M. K. (2007). Document reranking by term distribution and maximal marginal relevance for Chinese information retrieval. *Information Processing & Management*, 43(2), 315–326. <https://doi.org/10.1016/j.ipm.2006.07.011>

- Zarrinkalam, F., & Kahani, M. (2012). A new metric for measuring relatedness of scientific papers based on non-textual features. *Intelligent Information Management*, 4(4), 99–107. https://www.scirp.org/pdf/IIM20120400001_98298896.pdf
- Zhang, D., Luo, T., Wang, D., & Liu, R. (2015). *Learning from LDA using deep neural networks*. arXiv. <https://arxiv.org/pdf/1508.01011.pdf>
- Zhang, J., Zeng, J., Yuan, M., Rao, W., & Yan, J. (2016). LDA revisited: Entropy, prior and convergence. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1763–1772). ACM. <https://doi.org/10.1145/2983323.2983794>
- Zhou, H. K., Yu, H. M., & Hu, R. (2017). Topic discovery and evolution in scientific literature based on content and citations. *Frontiers of Information Technology & Electronic Engineering*, 18(10), 1511–1524. <https://doi.org/10.1631/FITEE.1601125>
- Zou, L., Liu, X., Buntine, W., & Liu, Y. (2021). Citation context-based topic models: Discovering cited and citing topics from full text. *Library Hi Tech*, 39(4), 1063–1083. <https://doi.org/10.1108/LHT-01-2021-0041>

How to cite this article: Akbulut, M., & Tonta, Y. (2026). Incremental refinement of relevance rankings: Balancing relevance depth and scope. *Journal of the Association for Information Science and Technology*, 77(10), 1247–1270. <https://doi.org/10.1002/asi.70101>