Received: 5 December 2023 | Revised: 6 April 2024 | Accepted: 10 April 2024

DOI: 10.1002/widm.1541

SYSTEMATIC REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

A systematic review of multidimensional relevance estimation in information retrieval

Georgios Peikos | **Gabriella Pasi**

Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

Correspondence
Georgios Peikos and Gabriella Pasi, Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milan I-20126 MI, Italy.
Email: g.peikos@campus.unimib.it and gabriella.pasi@unimib.it

Funding information
European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie, Grant/Award Number: 860721

Edited by: Elisa Bertino, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract

In information retrieval, relevance is perceived as a multidimensional and dynamic concept influenced by user, task, and domain factors. Relying on this perspective, researchers have introduced multidimensional relevance models addressing diverse search tasks across numerous knowledge domains. Through our systematic review of 72 studies, we categorize research based on domain specificity and the distinct relevance aspects employed for estimating multidimensional relevance. Moreover, we highlight the approaches used to aggregate scores related to these factors and rank information items. Our insights underline the importance of concise definitions and unified methods for estimating relevance factors within and across domains. Finally, we identify benchmark collections for evaluations based on multiple relevance aspects while underscoring the necessity for new ones. Our findings suggest that large language models hold considerable promise for shaping future research in this field, mainly due to their relevance labeling abilities.

This article is categorized under:
Application Areas > Science and Technology
Technologies > Computational Intelligence

KEYWORDS

information retrieval, multiaspect relevance, multidimensional relevance, systematic review

1 | INTRODUCTION

Users' engagement in search activities is commonly motivated by tasks stemming from persistent and evolving problematic situations (Belkin, 2016). In such searches, as users are presented with information items provided by search systems, a complex cognitive decision-making process is initiated, ultimately leading to them choosing useful items for further examination. The decision-making process is grounded in what Vickery (1959a) termed *user relevance*, which is commonly acknowledged in the field of information retrieval (IR). This notion of relevance concerns how users evaluate information as pertinent to their information needs. Changing from a user-centric perspective to a system-oriented one, we focus on the inherent mechanisms by which search systems operate. Central to their operation is a concept highlighted by scholars, as *system relevance* (Saracevic, 2016; Vickery, 1959b). This concept encapsulates a system's ability to retrieve information items in line with an information need and, consequently, estimate their relevance based on an algorithm or model. System relevance serves as an approximation to the aforementioned user relevance, aiming to align system outputs with user expectations.

IR systems commonly approximate user relevance by estimating how closely the content of documents aligns with the textual content of the expressed information need (i.e., query). This form of relevance is termed as *topical relevance*, and in numerous studies it is simply referred to as relevance. Another direction of research investigates the dynamic nature of relevance, that is, how the perception of relevance changes over time and through user–system interactions, as seen in studies focused on interactive IR (Liu, 2021). Other researchers investigate the multidimensional notion of relevance, suggesting that it is shaped by factors related to the user, the undertaken task, and the knowledge domain. Finally, all notions of relevance can be considered holistically, in systems using relevance models that rely on multiple factors and account for their evolving nature during search.

In this systematic review, we focus on studies that estimate multidimensional relevance, leveraging factors such as topicality, credibility, and timeliness of information, among others. Some of the proposed relevance models also acknowledge the dynamic nature of relevance, considering that the importance of the identified factors may change. The foundational premise of such studies is that by integrating multiple factors or considering their evolution over time, relevance models can more accurately approximate the user perceived relevance in certain search situations.

Apart from studies that propose multidimensional relevance models, the literature identifies two complementary research areas. The first is research that conceptualizes the notion of relevance in IR, across distinct knowledge domains and diverse search tasks. The second involves research centered on user studies, intending to discover the essential factors influencing user relevance within specific domains or search tasks. We briefly discuss such studies in Section 2, as they fall outside the main scope of this review but provide useful insights.

Recent technological advancements have expanded the horizons for developing novel multidimensional relevance models. A standout advancement in this regard is the rise of large language models (LLMs), distinguished by their advanced text generation capabilities. These models can improve the estimation of topical relevance and also have the potential to serve as effective tools for obtaining relevance judgments (Faggioli et al., 2023; Frei & Kramer, 2023). Consequently, the shift from primarily topically oriented relevance models to multidimensional relevance models is increasingly within reach.

Our survey systematically examines 72 studies that have proposed and experimentally evaluated multidimensional relevance models. We synthesize these studies based on their knowledge domains and search tasks, their employed relevance factors, and the utilized benchmark collections. The specific research questions tailored to our study are presented in Section 3. This review aims to aid the development of future multidimensional models by identifying current necessities and paving the way for future research.

The article is structured as follows. In Section 2, we introduce and briefly comment on two research areas associated with the notion of relevance in the field of IR. Moving forward, Section 3 presents the research aim and questions that guide our investigation. In Section 4, we present the outcomes of our synthesis addressing the posed research questions. Section 4.1 highlights the knowledge domains in which relevance has been perceived and modeled based on multiple factors. Section 4.2 analyzes the identified relevance factors based on their definitions and operationalizations, aiming to highlight their commonalities and differences across and within domains. Section 4.3 discusses the benchmark collections used to evaluate multidimensional relevance models in the included studies. Section 5 offers an in-depth discussion of our systematic literature examination findings, pointing to potential avenues for future research directions. Section 6 discusses our study's prospects and limitations, while Section 7 concludes our study. Finally, Appendices A and B provide details related to the systematic methodology employed for collecting and synthesizing literature studies and presents their overall characteristics.

2 | THE NOTION OF MULTIDIMENSIONAL RELEVANCE IN IR

From the beginning of information systems in the late 1950s to the present, the notion of relevance in IR has been the focal point of numerous research studies. At its core, relevance always indicates a relation (Saracevic, 2016). A fundamental distinction can be observed in the dual nature of relevance: user relevance, which captures the individual's perception of information, and system relevance, which is algorithmically determined by a system (Cooper, 1971; Swanson, 1986; Vickery, 1959a, 1959b). Over time, scholars from varied backgrounds have proposed different definitions to capture the notion of relevance. These definitions range from *affective relevance* tied to users' emotions and motivations, to *situational relevance* addressing specific tasks, *system or algorithmic relevance* determined by query and information matching using an algorithm, *topical relevance* focusing on the relation between the topic expressed in a query and topic covered by information objects, and *cognitive relevance* connecting a user's knowledge and the

information's novelty (Belkin, 2016; Borlund, 2003; Cosijn, 2009; Cosijn & Ingwersen, 2000; Ingwersen & Järvelin, 2005; Mizzaro, 1998; Saracevic, 1997). While each definition adopts a distinct viewpoint, they all describe a form of relationship to information. We direct readers interested in a comprehensive understanding of relevance in information sciences to the book by Saracevic (2016). Although the book offers an in-depth exploration of the topic, it does not review the technological and algorithmic perspectives. Specifically, it does not discuss how systems estimate multidimensional relevance, which is the main focus of our review.

Relevance has also been investigated within certain domains, as researchers have attempted to decompose it and identify the factors that contribute to information's usefulness for users. van Opijnen and Santos (2017) conduct a thorough examination of the notion of relevance within the legal domain, drawing upon the manifestations of relevance as outlined by Saracevic (2016). The notion of relevance has also been conceptualized in the e-commerce domain. Tsagkias et al. (2021) conclude that e-commerce relevance is determined by four primary dimensions, namely, user, time, query, and context (i.e., a product's category), emphasizing its domain-specific nature. Building on the framework introduced by Mizzaro (1998), Crestani et al. (2017) specify the notion of relevance in mobile searches. Furthermore, Balagopalan et al. (2023) explore the role of relevance in ensuring fair ranking.

Mainly by conducting user studies, numerous scholars have identified factors (i.e., *relevance factors*) that users take into account when assessing relevance in specific search scenarios, that is, by investigating what is referred to as *user relevance*. While a comprehensive examination of all these studies is beyond the scope of our review, we highlight a few representative ones here. For a more extensive exploration, readers can refer to the book by Saracevic (2016), as a starting point. Some key studies in this research field are the studies by Cool et al. (1993), Barry and Schamber (1998), and Xu and Chen (2006), among others. Xu and Chen (2006) conduct a user study centered on web searches. They investigate the significance of factors such as scope, novelty, topicality, reliability, and understandability in these searches. The findings highlight that topicality and novelty are the foremost factors for relevance, with understandability being the subsequent priority. Sun et al. (2019) in their systematic literature review identify the factors and indicators consumers use to evaluate the quality of online health information. Their research highlights trustworthiness, expertise, and objectivity as the most important across studies. In addition, dominant indicators are related to the web page's source, content, and design. Other studies reveal that assessing relevance based only on *topicality* is not sufficient for medical experts, as they leverage their own knowledge and experience (Tamine & Chouquet, 2017). Similar studies can be also found in other domains, such as the legal domain. The study by Wiggers et al. (2018) identifies factors affecting relevance assessment in legal professional searches, such as document type, recency, depth level, and legal hierarchy. Also Chu (2011) aims to discern factors influencing relevance judgments and their relative significance in legal search. The study highlights several relevance factors, with specificity/amount of information, ease of use, and subject matter are the most essential. The findings from the aforementioned and other related studies hold significant value. Mainly because they can guide the development of retrieval systems specifically tailored to certain search situations, ensuring a better approximation to user relevance in these tasks.

Drawing from the studies and definitions mentioned above and also supported by the study of Schamber et al. (1990), the notion of relevance emerges as a *multidimensional cognitive concept* influenced by users' perceptions of information and their distinct contextual situations (e.g., search task, domain). This concept is also *dynamic*, depending on users' perspective of the provided information in time. Nonetheless, as Schamber et al. (1990) conclude, relevance is a complex but *systematic and measurable* concept. In our review, we define multidimensional relevance as the concept of estimating relevance in IR systems (i.e., a type of *algorithmic relevance*), by considering multiple relevance factors. These factors influence relevance estimation in specific search tasks and can be related to user or domain characteristics, or the nature of the search task (e.g., professional search task). Consequently, multidimensional relevance systems aim to integrate multiple relevance factors in the retrieval process to approximate *user relevance*.

3 | RESEARCH AIM AND QUESTIONS

The main objective of this systematic review is the analysis of studies that consider relevance as a multidimensional notion and propose models and systems for its estimation. We categorize the identified studies based on their applied knowledge domain (e.g., health, legal, academic) and the relevance factors they utilize. In addition, we analyze the methods employed to aggregate these relevance factors. Furthermore, we group the different relevance factors used in the reviewed studies according to their definitions and operationalization, that is, how the authors estimated or measured these factors. We compile a comprehensive list of benchmark collections that have been utilized in the reviewed

studies. These benchmark collections are characterized based on the annotated relevance factors, the knowledge domain, their size, and availability. Finally, we provide an overview of various initiatives that offer shared tasks centered around multidimensional relevance. The ultimate goal of this systematic review is to shed light on the multidimensional nature of relevance and to highlight the various approaches and benchmark collections used to study this important concept across different knowledge domains. By doing so, we aim to contribute to a clearer understanding of multidimensional relevance and its practical and theoretical implications.

Following the methodological approach proposed by Cooper et al. (2019), the systematic review conducted in this study consists of the following steps: (1) Formulation of the research questions; (2) establishment and clarification of the inclusion and exclusion criteria associated with the selection of research papers; (3) development of a retrieval strategy (e.g., involved sources and databases, keywords); (4) proposal of a coding scheme for paper annotation; (5) synthesizing the findings to answer the research questions. This section introduces the research questions that guide our systematic review, while the remaining steps (2–5) are analyzed in Appendix A.

3.1 | Step 1: Research questions

By addressing these questions, we aim to gain valuable insights into how relevance is perceived, decomposed into several factors, and estimated in different knowledge domains. The answers to these questions will not only deepen our understanding of multidimensional relevance but also contribute to the advancement of research and its practical applications. To this end, this systematic review seeks to answer the following research questions:

RQ1. How is relevance conceptualized and operationalized as a multidimensional concept (as defined in Section 2) in the identified studies?

1. What are the different knowledge domains (e.g., health, legal, academic) in which multidimensional relevance has been explored?
2. What are the relevance factors utilized by researchers in the reviewed studies?
3. What are the diverse approaches employed to aggregate relevance factors in the context of multidimensional relevance estimation?

RQ2. How do authors defined and operationalized relevance factors (i.e., estimate a score to be associated with them) in the reviewed studies?

1. How have the relevance factors been defined within the studies incorporated in the review?
2. What methodologies are used to operationalize the identified relevance factors?

RQ3. Which benchmark collections have been used to estimate multidimensional relevance, and how are they characterized based on their annotated relevance factors, size, and availability?

The complete list of our study's inclusion and exclusion criteria is presented in Table A1 in Appendix A. Nonetheless, it is worth-mentioning here that our review exclusively included studies focusing on text retrieval systems (i.e., document retrieval), as studies involving other types of information objects (e.g., audio, video) would significantly expand the scope of the review.

4 | RESULTS

In this section, we analyze and synthesize the publications in our review, following the established coding scheme to answer the posed research questions. We outline that the included publications have been analyzed based on their characteristics, specifically their geographical distribution, the collaborative efforts between industry and academia highlighting synergies, the diversity in types of publication venues, and the temporal distribution that offers insights into the evolution of research in this domain. This analysis is presented in Appendix B.

4.1 | How is relevance conceptualized and operationalized as a multidimensional concept?

In this section, we provide insights related to the conceptualization of relevance across different knowledge domains. Our primary goal is to identify and describe these domains, and highlight particular search tasks in which relevance has been treated as a multidimensional notion. Following that we mention the specific relevance factors that are utilized within each domain and search task. Finally, we classify the various methods that researchers have used to combine relevance scores associated with distinct factors, to obtain an overall multidimensional relevance score.

4.1.1 | What are the different knowledge domains in which a multidimensional notion of relevance has been explored?

The left column of Table 1 presents a detailed breakdown of studies conducted across diverse knowledge domains and search tasks, from which we have identified 18 domains. The observed domains span from academic and medical to web and social, with some emphasizing specific search tasks, like consumer health and biomedical article retrieval tasks. Notably, while some domains have only 1 study, research areas like web search dominate with 25 studies, reflecting possible research emphasis and potential complexity of investigating multidimensional relevance in the other domains. Further result analysis, presented in Section 4.3, deepens our comprehension of the underlying reasons for the observed long-tailed distribution of the identified domains.

Having established the distribution of various knowledge domains, here we focus on each domain and highlight the specific retrieval tasks in the identified studies. In web search, Lioma et al. (2016) explore how the factuality and objectivity of documents relate to document relevance, and integrate them as query-independent features in a retrieval model. Undoubtedly, PageRank is a fundamental feature integrated into commercial web search systems (Brin & Page, 1998). Expanding on that, Craswell et al. (2005) implement sigmoid transformations on PageRank, URL Length, and ClickDistance and combine them with topicality scores such as BM25. Other scholars explore how external knowledge from knowledge graphs can be combined with topicality to improve retrieval performance (Li et al., 2021; Rinaldi, 2009). Focusing on specific web search tasks, several studies propose retrieval models that integrate topicality with other relevance factors such as information freshness (Bambia & Faiz, 2015; Dai et al., 2011), content's quality (Bendersky et al., 2011), content's readability (Sasaki et al., 2016), source's popularity, recency, and reputation (Badache & Boughanem, 2014). Other scholars proposed models to retrieve child-friendly content (Eickhoff et al., 2013), information related to programming search tasks (Silva et al., 2019), and web tables (Shraga et al., 2020). Several studies leverage user-related relevance factors for web retrieval, that is, personalized web search (Collins-Thompson et al., 2011; Li et al., 2017; Sahraoui & Faiz, 2017; Sieg et al., 2007; Uprety et al., 2018). Moreover, research efforts have been made to tackle the challenge of obtaining a diverse set of retrieved documents, ensuring they address multiple query aspects while reducing redundancy (topic distillation) (Farah & Vanderpooten, 2008; Shajalal & Aono, 2020; van Doorn et al., 2016; Vargas et al., 2012). Finally, several scholars propose frameworks that leverage multiple relevance factors for document ranking and use web search as an application domain (Eickhoff & de Vries, 2014; Komatsuda et al., 2016; Zhuang et al., 2021).

Within the medical domain, two distinct search tasks where relevance is interpreted as a multidimensional concept have been identified: the retrieval of biomedical articles (Alsulmi & Carterette, 2018; Qu et al., 2020, 2021; Xu et al., 2016; Znaidi et al., 2016) and consumer health search (Palotti et al., 2019; Putri et al., 2021; Upadhyay et al., 2023; van Doorn et al., 2016; Zhang et al., 2015). In addition, research endeavors prioritize retrieving health information that is topically relevant, credible, reliable, and correct (Banerjee et al., 2023; Fernández-Pichel et al., 2022; Upadhyay et al., 2022). Additional domains that have attracted the attention of researchers with respect to multidimensional relevance estimation include social and e-commerce searches. In social search, studies explore Twitter (now referred to as X Corp) search and integrate topical relevance with signals like recency, authority, trustworthiness (Jabeur et al., 2012; Moulahi, Moulahi, et al., 2014; Ravikumar et al., 2013). Other studies focus on retrieving content related to events, disasters, or opinions (Madisetty & Desarkar, 2022; Putri et al., 2020), or leverage social content to improve ranking (Tamine et al., 2011). E-commerce has risen to significant prominence in recent years. In this domain, the notion of relevance is influenced by domain-specific factors that are related to products, temporal contextual information (referred to as *seasonality*), reviews, and users' intents, among others (Bassani & Pasi, 2021; Carmel et al., 2020; Feng et al., 2018; Karmaker Santu et al., 2017; Mandayam Comar & Sengamedu, 2017; Yang et al., 2021).

TABLE 1 representing the various knowledge domains and search tasks alongside the exploited relevance factors for multidimensional relevance estimation.

Knowledge domain and search tasks	Relevance factors
Web search (total studies: 25)	Topicality, reputation, readability, PageRank, authority, objectivity, knowledge, content quality, popularity, freshness, factuality, coverage, anchor text, user's actions, temporal relevance, syntactic relevance, other task-based features
• Personalization ($N = 5$)	Topicality, user's interest, scope, reliability, user's habit, novelty
• Topic distillation ($N = 4$)	Topicality, rareness, proximity, prominence, position, frequency, document length, content diversity, authority
• Child-friendly content retrieval ($N = 1$)	Topicality, appropriateness for children
• Programming-related search ($N = 1$)	Topicality, semantic similarity, API method-based score, API class-based score
• Table retrieval ($N = 1$)	Topicality, multi-modal table properties
Medical search (total studies: 11)	Topicality, passage-level reliability, passage-level topicality, genuineness, correctness
• Biomedical articles search ($N = 5$)	Topicality, content diversity, other task-based relevance
• Consumer health search ($N = 5$)	Topicality, understandability, credibility, readability
Social search (total studies: 7)	
• Twitter search ($N = 3$)	Topicality, trustworthiness, temporal relevance, recency, authority, user's social importance
• Disaster-related search ($N = 1$)	Topicality, informativeness, interestingness, credibility, opinionatedness
• Event-related search ($N = 1$)	Topicality, hashtag-based similarity, event-based topicality
• Opinion-related search ($N = 1$)	Topicality, informativeness, interestingness, credibility, opinionatedness
• Scientific community search ($N = 1$)	Topicality, user-related social features, popularity, freshness
E-commerce search (total studies: 6)	Topicality, temporal relevance (*Seasonality*), sales, reviews, purchase user intent, node compatibility, item popularity, category compatibility, other task-based features
Academic search (total studies: 4)	Topicality, reliability, recency, readability, coverage
• Math search ($N = 1$)	Image similarity and context similarity based on math formulas
Blog post search (total studies: 4)	
• Opinions search ($N = 4$)	Topicality, topical evidence, temporal relevance, social features, opinion, authoritative evidence
Newswire stories search (total studies: 4)	Topicality, reliability, objectivity, freshness, coverage, user-related appropriateness, factuality
Community question answering (total studies: 2)	Topicality, recency, passage quality
Geographic information retrieval (total studies: 2)	Topicality, temporal relevance, spatial relevance
Legal search (total studies: 2)	Document's usage, citations, other task-based features
Educational search (total studies: 1)	Task-based features
Expert finding (total studies: 1)	
• Expert translator finding ($N = 1$)	Topicality (as a proxy to language proficiency), price, number of cooperation times, duration of the translation
Local search (total studies: 1)	Topicality, reputation, distance
Math search (total studies: 1)	Taxonomic distance of functions, data type hierarchical level, match-depth, coverage, other task-based features
Mobile search (total studies: 1)	Topicality, location, user's interest
Personalized bookmark search (total studies: 1)	Topicality, user-based relevance

TABLE 1 (Continued)

Knowledge domain and search tasks	Relevance factors
Personalized contextual search (total studies: 1)	User's interest, location
XML retrieval (total studies: 1)	Topicality, specificity, exhaustivity

Research on multidimensional relevance estimation spans a variety of other domains, reflecting the diverse nature of information needs across different contexts. For example, in academic search, researchers such as Jomsri and Prangchumpol (2015), Arastoopoor (2018), and Singh and Dave (2019) have put forth models incorporating recency alongside other domain-specific factors. Meanwhile, math search is another domain where the complexity of relevance estimation necessitates the combination of multiple signals, as shown by Yan et al. (2022). Blog post search involves the aggregation of signals related to a source authority or level of opinion (Chenlo et al., 2015; Eickhoff et al., 2013; Gerani et al., 2012; Huang et al., 2018). In newswire search, researchers have proposed models that leverage recency, reliability and coverage signals (da Costa Pereira et al., 2009, 2012; Dumitrescu & Santini, 2021; Lioma et al., 2016). Geographic IR is distinguished by its integration of temporal, spatial, and topical relevance that are commonly used in the domain (Daoud et al., 2013; Palacio et al., 2010). In community question answering topical relevance mainly refers to text passages and is combined with factors like recency and context's quality (Amancio et al., 2021; Yulianti et al., 2018). Another identified domain is referred to as educational search in which primary school children are considered as users (Usta et al., 2021). Legal search has witnessed recent explorations, as reflected by the included studies (Ma et al., 2023; Wiggers et al., 2023), while, in this domain, the conceptualization of relevance significantly diverges from other domains, as we analyzed in Section 2. Additional domains include expert finding, with specific areas like expert translator finding (Rekabsaz & Lupu, 2014), local search (Kang et al., 2012), mobile Search (Bouidghaghen et al., 2011), personalized bookmark search (Eickhoff et al., 2013), personalized contextual search (Moulahi, Tamine, & Yahia, 2014), and XML Retrieval (Ashoori & Lalmas, 2007).

Our analysis reveals the widespread adoption of multidimensional relevance models across diverse knowledge domains and search tasks. The identified synergy between academia and industry underscores a close relationship between real-world applications and ongoing research advancements in this field.

4.1.2 | What are the relevance factors utilized by researchers in the reviewed studies?

As it can been seen in Table 1, each domain has its unique set of relevance factors, some of which are shared across domains. This showcases the multidimensional and task-specific nature of IR across diverse domains and search tasks. The analysis of the included studies revealed that certain domains are dominated by identical relevance factors; for instance, medical searches are often influenced by factors associated with the credibility of the information. Furthermore, some relevance factors remain consistent across multiple domains, exemplified by the usage of the recency factor regardless of the domain or task. Notably, there are relevance factors that essentially convey similar relevance signals but are mentioned differently, underscoring the need for future formalization to bring consistency. This is seen in terms such as credibility, trustworthiness, and genuineness, which although distinct in wording, often intersect in their conveyed meaning. In Section 4.2, addressing our second research question, we aim at analyzing relevance factors that fall in the aforementioned category by analyzing their definitions and operationalization.

In Table 1, we reference the terms *Task-based Features* and *Task-based Relevance*. Recognizing that these terms might hold varying interpretations, we highlight their meaning within the framework of our review. We use the term *Task-based Features* when a study incorporates a considerable volume of features to estimate multidimensional relevance, often in a learning to rank (LtR) setting. This was evident in two studies in the e-commerce domain. In the study by Karmaker Santu et al. (2017), a set of 562 features is utilized, focusing on aspects related to the query, the document (in this case, a product), and the query-document relationship. These features encompass metrics such as BM25 scores, user ratings, and total sales. Similarly, Feng et al. (2018) deploy a variety of features to determine relevance. While the exact number of these features is unspecified, some illustrative examples include the item's popularity and rating score. Moving to the educational search domain, Usta et al. (2021) leverage 50 domain-specific and generic features. These

related to queries (e.g., the name of a course), documents (for instance, the document's course), their relationship (like BM25), and they leverage session data. In legal search, Ma et al. (2023) also generate a set of domain-specific features. Specifically, the authors, leveraging the structure of legal documents, they split them in three core segments, namely *Facts*, *Holding*, and *Decision*. By doing that, they create a token-level representation for each of the segments, concatenate them, and use them to train a LtR model. Similarly, in web search, Zhuang et al. (2021) propose the use of generalized additive models (GAMs) for ranking, in an approach that also leverages a vast amount of domain-specific and generic features. Lastly, in the medical search, Alsulmi and Carterette (2018) leverage 74 features for biomedical articles search. Regarding the term *Task-based Relevance*, this is used to describe three studies from biomedical articles search (Qu et al., 2020, 2021; Znaidi et al., 2016). In these studies, the authors model relevance estimation by considering several relevance factors, and the characteristics of the search tasks. Specifically, the authors propose approaches that mimic the user's workflow and decision-making processes and develop search models that follow the same steps to predict a document's relevance.

A more detailed analysis of the identified relevance factors can be found in Section 4.2, where we discuss proposed definitions and operationalization methods.

4.1.3 | What are the diverse approaches employed to aggregate relevance factors in the context of multidimensional relevance estimation?

In this section, we focus on the methodologies the authors adopt to aggregate multiple relevance factors into a unified relevance score. We categorize these methodologies as *data driven*, *model driven*, and *other* that includes studies that do not fall in either of these categories. Data-driven methods primarily use LtR or other machine learning techniques. Model-driven methods have been employed in the majority of the reviewed studies. Notably, the most frequent approach is to perform a linear combination of the consider relevance factors. While our review primarily explores multidimensional relevance models, we acknowledge studies that leverage score fusion techniques, as it is a popular method for aggregating scores from distinct relevance factors. Figure 1 presents the distribution of studies based on their aggregation approach types, indicating that 40 studies employ a model-driven approach, 24 adopt a data-driven approach, and the remaining utilize result fusion, other methods, or do not specify their aggregation technique.

Model-driven approaches. With a few exceptions, the majority of the model-driven approaches exploit a weighted linear combination to obtain an overall relevance score. Nonetheless, some exceptions do exist. *Linear combination* (or weighted linear combination) has been exploited to aggregate scores related to distinct relevance factors in web search (Craswell et al., 2005; Lioma et al., 2016; Rinaldi, 2009; Sahraoui & Faiz, 2017; Silva et al., 2019), math search (Zhang & Youssef, 2014), academic search (Jomsri & Prangchumpol, 2015; Singh & Dave, 2019), blog post search (Chenlo et al., 2015; Gerani et al., 2012; Huang et al., 2018), medical search (Banerjee et al., 2023; Upadhyay

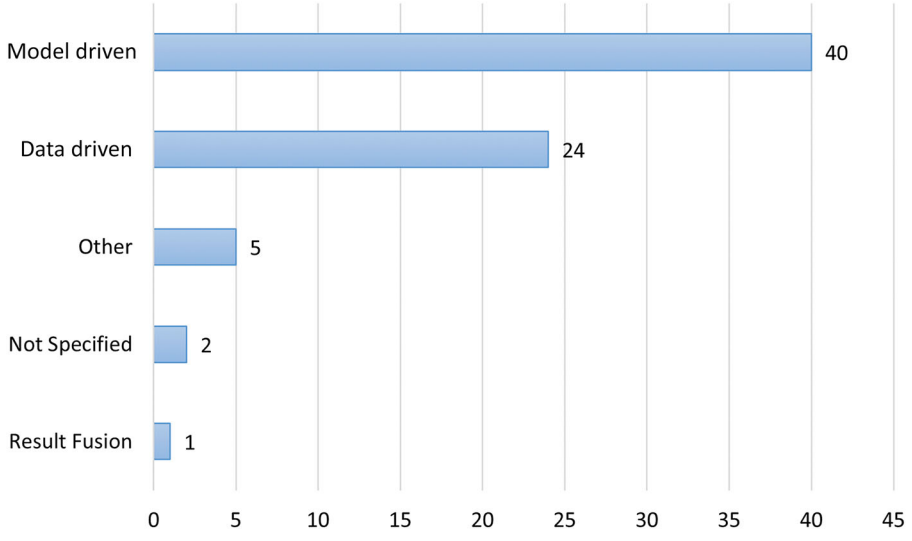

FIGURE 1 Number of studies categorized based on the employed aggregation approach.

et al., 2022), social search (Madisetty & Desarkar, 2022; Putri et al., 2020; Tamine et al., 2011), e-commerce (Bassani & Pasi, 2021), community question answering (Yulianti et al., 2018), and geographic IR (Daoud et al., 2013). In consumer health search, Zhang et al. (2015) introduce a custom formula that applies an exponential weighting to the readability score and then multiplies it with a power-weighted topical relevance score. Another popular model-driven aggregation technique relies on *Copulas*. Copulas is a class of probability density functions that can be used to describe the dependence between multiple variables, separate from their individual behaviors or distributions. They excel at capturing complex, nonlinear relationships, including the intricate connections seen at extreme values, known as tail dependencies. Due to that, several studies in our review leverage copulas for multidimensional relevance estimation (Eickhoff et al., 2013; Eickhoff & de Vries, 2014; Komatsuda et al., 2016; Sasaki et al., 2016; Sieg et al., 2007). In their study, da Costa Pereira et al. (2009) propose the usage of a *prioritized scoring aggregating* operator for multidimensional relevance estimation that assumes order of importance among the relevance factors. In detail, the importance weight of a certain criterion (i.e., relevance factor) is dependent upon the satisfaction or score of a previous or higher-priority criterion. Bouidghaghen et al. (2011) introduce another operator for multidimensional relevance estimation, namely the *prioritized* "and" *operator*. The distinguishing aspect of this operator is the extent to which the least satisfied criterion is considered. Since their introduction, these operators have been used in several studies (da Costa Pereira et al., 2012; Znaidi et al., 2016). In math search, Yan et al. (2022) leverage the *hesitation fuzzy set* to obtained an interpretable document ranking. Other scholars have modified traditional *language models* by incorporating additional relevance factors. Specifically, they integrated these factors as prior probabilities or made specific adjustments to existing models (Ashoori & Lalmas, 2007; Badache & Boughanem, 2014; Bambia & Faiz, 2015). Other studies introduce *relevance models*, for example, probabilistic models, that account for several relevance aspects, based on the characteristics of the applied domains (Bendersky et al., 2011; Jabeur et al., 2012; Uprety et al., 2018; Vargas et al., 2012). Finally, some studies consider the task of multidimensional relevance estimation as a multicriteria decision-making (MCDM) problem (Farah & Vanderpooten, 2008; Moulahi, Moulahi, et al., 2014). Therefore, these studies leverage MCDM methods such as the *Choquet Integral* or *ELECTRE*. It is worth noting that several studies mentioned above exploit some training data to predict a set of importance weights associated with the relevance factors.

Data-driven approaches. For the data-driven approaches, their diversity makes it challenging to identify commonalities and categorize them; thus, we provide a concise description of each study. Numerous data-driven techniques have emerged in the field of e-commerce. Mandayam Comar and Sengamedu (2017) leverage users' search intents and propose a multi-intent Poisson-beta model for product ranking. The model, which identifies users' purchase intentions based on observed click patterns, is trained using click logs data collected over 30 days from the Amazon product search dataset. Karmaker Santu et al. (2017) experimented with several LtR methods and found LambdaMART as the best performing for product search. The authors emphasize the efficacy of popularity-based features and found that click rates are more predictable than add-to-cart ratios. Their experimentation shows that model optimization based on order rates frequently yields the most consistent predictions, indicating a potential advantage in transitioning to order rate-centric models. Feng et al. (2018) propose the Multi-Agent Recurrent Deterministic Policy Gradient tailored for multi-scenario ranking in the e-commerce domain. The model uses an online learning system that dynamically updates based on real-time user logs and a replay buffer mechanism. Consequently, it can continuously adapt to changing user behaviors. The study by Li et al. (2021) presents the topic-enhanced knowledge-aware retrieval model, which incorporates three dimensions of relevance, that is, semantic similarity, knowledge relevance, and topical relatedness, to assess the relevance between a query and a document. The model aims to minimize simultaneously a ranking loss that ensures good semantic relevance, and the loss of the neural model that ensures topical relatedness. Yang et al. (2021) introduce LogSR and VelSR features based on neural models to capture product seasonality in e-commerce search. They incorporated these features into a standard LtR setup, validated their approach through offline and online experiments, and highlighted its efficacy. Finally, Carmel et al. (2020) address the challenge of optimizing multiple objectives, including maximizing product relevance and purchase likelihood simultaneously (a problem known as Multi-Objective Ranking Optimization [MORO]). To that aim, the authors introduce a novel approach, namely stochastic label aggregation. This method randomly assigned labels to training examples based on a given distribution over the labels. Theoretical analysis and empirical experiments on different datasets revealed that MORO with stochastic label aggregation consistently outperformed deterministic label aggregation methods. Label aggregation has also been exploited as an approach in local search by Kang et al. (2012). The authors define a label aggregation function that quantitatively combines multi-aspect relevance values into an overall score. To train this function, they use relative preference data, where one document is preferred over another. Once the aggregation function is learned, it is applied to a larger dataset containing ranking features and multiaspect relevance vectors. This process generates an expanded dataset with overall relevance

scores. Subsequently, they train a ranking function using this expanded dataset, enabling effective handling of multiple relevance dimensions in ranking models. In their study, van Doorn et al. (2016) perceive multiple relevance factors as objectives and aim to learn a set of rankers that provide different trade-offs concerning these objectives. They use a combination of gain-based evaluation and multiobjective optimization techniques, including optimistic linear support and dueling bandit gradient descent, to find optimal rankers. In medical search, several LtR and machine learning approaches have been introduced, with few emphasizing interpretability. Applying LtR techniques for retrieving biomedical articles, Alsulmi and Carterette (2018) exploit a wide range of general and domain-specific features for ranking. Notably, among the algorithms investigated in the research, Coordinate Ascent emerged as the top-performing when combined with a feature selection strategy. For biomedical article retrieval, Xu et al. (2016) introduce a framework that combines multiple LtR techniques. This framework aims to optimize document ranking by considering topical relevance and diversity. The authors utilized label aggregation approaches to merge these two aspects and train the LtR models. Among all of the evaluated models, LambdaMART exhibited the best performance. Qu et al. (2020, 2021) propose a model that leverages structured search strategies to build an effective, explainable, and label-efficient retrieval algorithm for professional search tasks. This model utilizes machine learning classifiers to predict different aspects of the query and then combines these predictions using a logical function to determine document relevance. The experimental results show that their model performs as well as complex LtR models, even with limited labeled documents. In consumer health search, Putri et al. (2021) introduce a multitask learning model that simultaneously estimates relevance based on topicality and another factor, such as readability or credibility. Their approach combines a neural retrieval model for topical relevance estimation with a classification model that categorizes documents based on the aforementioned factors. Both of these models share certain model parameters during training and inference. Also, in the context of consumer health search, Palotti et al. (2019) explore various methods to incorporate understandability and topicality into ranking. Of the tested methods, the authors concluded that LtR is the most effective. Fernández-Pichel et al. (2022) also leverage an LtR approach to rank health-related documents by considering several factors. Their experiments revealed that result fusion methods, such as CombSUM, outperformed LtR in terms of effectiveness. The web search domain has also witnessed the advent of various data-driven techniques. The study by Li et al. (2017) stands out as the most exhaustive one regarding the utilization of relevance factors and the depth of their feature engineering efforts. In the context of web search, the authors identify seven relevance factors, operationalize them using multiple features, and incorporate these features into an LtR model, that is, LambdaMART. In their work, Zhuang et al. (2021) present interpretable ranking models that utilize GAMs. These models can integrate both list-level and item-level features, making them well-suited for LtR tasks. In the context of web search, their experiments show that the proposed ranking GAMs outperform conventional GAMs while preserving their interpretability. Dai et al. (2011) introduce a criteria-sensitive divide-and-conquer ranking framework, an LtR methodology that optimizes topical relevance and freshness. The approach enhances the divide-and-conquer ranking technique by using hybrid labels and leveraging a new query-document importance factor that the authors introduced. Also, in web search, Collins-Thompson et al. (2011) propose an LtR method to re-rank topically relevant web pages according to their reading level. That is achieved by estimating the reading proficiency of users and the complexity of web pages, and by training a LambdaMART ranking model. Shraga et al. (2020) create a deep-learning retrieval technique for web table retrieval that considers web tables as multimodal entities. Their neural ranking model leverages gated multimodal units to represent queries and table modalities jointly. Experiments indicate the potential of viewing web tables as multimodal structures in future research. In expert finding, Rekabsaz and Lupu (2014) develop a translator-expert retrieval system that leverages domain-specific features such as price and delivery time, among others, for ranking. Through empirical evaluations, they determined that a ranking model based on linear regression leads to superior performance. Amancio et al. (2021) introduce a ranking approach for community question answering, leveraging quality and recency features. The authors experiment with nine LtR algorithms, from which Coordinate Ascent and LambdaMart lead to the best performance. Usta et al. (2021) employ an LtR approach specifically tailored for educational search. The model exploits features related to queries, documents, user's session, and their relationships. Furthermore, instead of using a single general model to rank all queries, the authors introduced query-dependent ranking models, grouping queries based on common characteristics, like association with a course user's grade level. These models lead to significant performance improvements. In legal search, Ma et al. (2023) propose a structured LtR model to retrieve the most relevant legal cases for a given query. Their method uniquely combines semantic-level and charge-level relevance scores by integrating internal case details with external structural information about charges. Utilizing the Lightgbm model, they effectively aggregate these scores to produce a ranked list of cases, using nDCG as a training objective.

Other approaches. The studies discussed below estimate multidimensional relevance by considering various relevance factors depending on the specific search tasks they address. In the context of academic and social search, the works by Arastoopoor (2018) and Ravikumar et al. (2013) both propose a retrieval pipeline that re-ranks an initial set of documents based on the considered relevance factor(s). Upadhyay et al. (2023) introduce a re-ranking approach that leverages a transformer-based model (i.e., BioBERT) for topically relevant and credible passage retrieval. Similarly, Shajalal and Aono (2020) sequentially re-ranks a set of documents, aiming to reduce information redundancy. Dumitrescu and Santini (2021) create a custom function highly tailored to the characteristics of the studied search task (i.e., newswire search). Lastly, Palacio et al. (2010), apply rank fusion techniques to combine relevance scores, while, due to limited information in the original studies, we cannot classify the methods exploited by Wiggers et al. (2023) and Brin and Page (1998).

4.2 | How do authors define and operationalize relevance factors (i.e., estimate a score to be associated with them) in the reviewed studies?

As highlighted in Section 4.1.2, some factors are recurrent across multiple domains, whereas others convey the same relevance signals but differ in terminology. To elucidate this, Sections 4.2.1 and 4.2.2 are dedicated to illustrating how the most frequent used relevance factors are defined and applied. To facilitate our analysis, in Table 2 we present a synthesis of these factors. On the left, it clusters similar relevance factors for easy comparison, and on the right, it enumerates the specific domains and search tasks where they have been operationalized.

4.2.1 | How have the relevance factors been defined within the studies incorporated in the review?

Examining Table 2 we address this research question by presenting and discussing the diverse definitions of the listed relevance factors.

Topicality. Across the included studies, following the standard paradigm in IR, *topicality* has been defined as the degree to which the content of a document matches or relates to a query posed by a user. Therefore, it is a factor that depends on the relationship between a query and a document.

Appropriateness, user's interest, personal relevance, appropriateness for children, user-related social features, user intent, user's habit. *Appropriateness* was introduced by da Costa Pereira et al. (2009) and later adopted in da Costa Pereira et al. (2012), both utilizing the same definition and operationalization. It has been defined as a relevance factor that estimates how appropriate a document is to the user's interest. The concept of *user's interest* has been referenced in multiple studies (Bouidghaghen et al., 2011; Dumitrescu & Santini, 2021; Li et al., 2017; Sahraoui & Faiz, 2017; Sieg et al., 2007; Tamine et al., 2011; Uprety et al., 2018). Yet, not every study provides a formal definition for it. Relying on previously introduced definitions, Tamine et al. (2011) consider that user interest expresses the cognitive background of the user. Li et al. (2017) define interest as the extent to which the user prefers the retrieved documents according to their topics of interest, whereas Uprety et al. (2018) adopt the same definition in their study. In their investigation, Bouidghaghen et al. (2011) utilize Park's (1994) definition, which assesses "Interest" as the degree to which a retrieved document aligns with the user's interest, a concept akin to appropriateness introduced by da Costa Pereira et al. (2009). *User's habit* has been defined by Li et al. (2017) as the extent to which the retrieved documents are preferred by a user according to their sources, genre, and language, among others. This definition has been adopted also by Uprety et al. (2018). Both the notions of *personal relevance* and *appropriateness for children* have been mentioned by Eickhoff et al. (2013). However, due to limited details in the original paper, it is challenging to provide their definitions in our review. Within the e-commerce domain, Mandayam Comar and Sengamedu (2017) identify and utilize two distinct *user intents*—purchase and explore—to rank products. The authors see purchase intent as akin to the navigation intents in standard web searches but with the user's goal directed toward finding a specific product. When users are curious to explore the variety of items displayed by the retrieval system, it is considered an exploration intent. Based on these definitions, these relevance factors calculate relevance signals that estimate the relationship between a user and a document.

Reliability, credibility, trustworthiness, genuineness, factuality and objectivity, correctness. The terms mentioned above have been used in multiple studies of our review and point to similar relevance signals. The notion of

TABLE 2 representing the most frequently used relevance factors across the included studies. The left column groups similar relevance factors together, while the right mentions the domains and search tasks in which they have been employed.

Relevance factors	Knowledge domain and search tasks
Topicality ($N = 64$)	Exploited in the vast majority of the included studies
Appropriateness ($N = 2$), user's interest ($N = 7$), personal relevance ($N = 1$), appropriateness for children ($N = 1$), users' intent ($N = 1$), user's habit ($N = 2$)	Web search (personalization, child-friendly content retrieval), social search (scientific community search), E-commerce, newswire stories search, mobile search, personalized bookmark search, personalized contextual search
Reliability ($N = 6$), credibility ($N = 2$), trustworthiness ($N = 1$), genuineness ($N = 1$), factuality and objectivity ($N = 1$), correctness ($N = 1$)	Web search (personalization), medical search (consumer health search), social search (Twitter, disaster-related and opinion-related searches), academic search, newswire stories search
Freshness ($N = 4$), temporal relevance ($N = 3$), recency ($N = 3$), novelty ($N = 2$)	Web search, social search (Twitter search, scientific community search), E-commerce search, academic search, blog post search, newswire stories search, community question answering, geographic information retrieval
Readability ($N = 6$), understandability ($N = 3$)	Web search, medical search (consumer health search)
Content diversity ($N = 4$), exhaustivity ($N = 1$), scope ($N = 2$)	Web search (topic distillation), medical search (biomedical articles search), XML retrieval
Authority ($N = 4$), authoritative evidence ($N = 1$)	Web search (topic distillation), social search, blog post search
Coverage ($N = 5$), specificity ($N = 1$)	Web search, academic search, newswire stories search, math search, XML retrieval
Spatial relevance ($N = 2$), location ($N = 3$)	Geographic information retrieval, local search, mobile search, personalized contextual search
Objectivity ($N = 1$), opinionatedness ($N = 1$), opinion ($N = 3$)	Web search, social search (disaster-related and opinion-related searches), blog post search
Content quality ($N = 1$), passage quality ($N = 2$), web page quality ($N = 2$)	Web search, community question answering
Popularity ($N = 2$), PageRank ($N = 2$), reputation ($N = 2$)	Web search, E-commerce search, local search

reliability in web search has been defined in the studies of Li et al. (2017) and Uprety et al. (2018) as the extent to which users trust a source, and it is associated with the wisdom of population. Similarly, in newswire stories search, da Costa Pereira et al. (2009, 2012) define reliability as the extent to which a user trusts a document's source, that is, a source's reputation. In a different direction, Fernández-Pichel et al. (2022) perceive the reliability of web content as a combination of content *correctness* and source credibility. Similarly, Banerjee et al. (2023) incorporate the concepts of credibility and correctness in their retrieval approach. The authors perceive correctness as a query-dependent relevance factor that assesses a document as correct if "it contains an answer that matches the topic's given answer." For further information regarding the notion of information correctness, as exploited by the authors, we refer interested readers to the overview paper of the Text Retrieval Conference (TREC) 2020 Health Misinformation Track (Clarke et al., 2020). In Twitter search, Ravikumar et al. (2013) employ the term *trustworthiness*, associating it with both the source and the content of a tweet. Putri et al. (2021), in consumer health search, acknowledge *credibility* and trustworthiness as two distinct notions which are mutually dependent. Upadhyay et al. (2022) introduce the concept of *genuineness* as a new abstract term that encompasses the various aspects introduced above (credibility, trustworthiness, among others). In their study, Lioma et al. (2016) use *factuality* and *objectivity*, estimated based on document contents, as proxies to estimate credibility. These relevance factors derive mostly from a document's attribute (i.e., its source) or its content. According to other researchers, these concepts are subject to users' perception on a document's source.

Freshness, temporal relevance, recency, novelty. Upon careful examination, the previously highlighted terms are indicative of relevance factors, which are designed to estimate comparable relevance signals (Amancio et al., 2021; Badache & Boughanem, 2014; Bambia & Faiz, 2015; Dai et al., 2011; Daoud et al., 2013; Dumitrescu & Santini, 2021; Jabeur et al., 2012; Jomsri & Prangchumpol, 2015; Li et al., 2017; Moulahi, Moulahi, et al., 2014; Omidvar-Tehrani et al., 2022; Yang et al., 2021). To enhance web search using social priors, Badache and Boughanem (2014) present a domain-specific interpretation of *freshness*, defining it as "a date of each social action (e.g., date of comment, date of

share) performed on a resource on social networks can be exploited to measure the recency of these social actions, hence freshness of information." Another study in web search aiming to answer real-time sensitive queries defines a document's freshness relying solely on its content and specifically by including "fresh words" (Bambia & Faiz, 2015). The authors consider as fresh words those that are trending on the social web and are topically relevant to the query, typically found in new social posts, micro-blogs, or breaking news. In web search, Dai et al. (2011) define freshness as a concept sensitive to the query temporal content, such as when users search for breaking news or events. In the context of newswire search and personalization, by considering also the notion of freshness, Dumitrescu and Santini (2021) argue that an item is considered fresh if it falls within a semantic domain of a user's interest that has not been encountered in the recent history. *Recency* is another term used in the literature. Amancio et al. (2021) conceptualize recency in community question answering by assessing the recency of the topics or terms present in an answer, that is, the answer's content. This definition aligns with the one given by Bambia and Faiz (2015). Lastly, Li et al. (2017) and Uprety et al. (2018) exploit the term *novelty*, drawing on the definition put forth by Xu and Chen (2006) who defined novelty as "the extent to which the content of a retrieved document is new to the user or different from what the user has known before" and argue that recentness can be regarded as one possible way of ensuring novelty, but not the only one. One can observe that even though these relevance factors share the same terminology (e.g., freshness) their definitions differ. Some scholars consider them dependent only on a document's attributes. In contrast, others describe the relationship between a document and a user or a query.

Readability, understandability. In their study, Sasaki et al. (2016) adopt the *readability* definition introduced by Klare (2000), in which "text readability can be formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the given material." The other studies in our review that utilize readability for document ranking do not mention a formal definition. Concerning *understandability*, both Li et al. (2017) and Uprety et al. (2018) treat the term as synonymous with readability. They adopt the definition from Xu and Chen (2006), which describe understandability as a "complex cognitive concept that measures the extent to which the user perceives the content of a retrieved document as easy to read and understand." In their work on consumer health search, Palotti et al. (2019) differentiate the notions of readability and understandability so that readability measures how easy it is to understand a text. Understandability is a broader term that encompasses the text's readability and presentation, such as its legibility, layout, and even the use of visuals to clarify complex ideas. Based on these definitions, readability and understandability depend on a document's characteristics (e.g., its presentation) and content. Nonetheless, these definitions imply that the concepts are subject to users' perceptions.

Content diversity, exhaustivity, scope. In biomedical article retrieval, Xu et al. (2016) incorporate *diversity* to maximize the coverage of query-related aspects in retrieved documents. Both Shajalal and Aono (2020) and Singh and Dave (2019) exploit information topicality and coverage, as described above, as a proxy to retrieve documents with diverse topics. Based on the studies mentioned before, one can observe a connection between coverage and topical diversification in the result list, where coverage serves as a means to attain diversification. In XML retrieval, Ashoori and Lalmas (2007) define *exhaustivity* based on the degree (i.e., how much) an XML element (i.e., a document's section) discusses the topic of the user's query. Li et al. (2017) and Uprety et al. (2018) leverage the notion of *scope* in their experiments. Relying on the definition of Xu and Chen (2006), the scope factor is defined as the extent to which the topic covered by a retrieved document is appropriate to the user's information need, that is, both breadth (similar to coverage/specificity) and depth (similarly to exhaustivity). Concluding, these relevance factors are defined based on the relationship between a query and documents.

Authority, authoritative evidence. In social search, Moulahi, Moulahi, et al. (2014) define authority as the influence of tweets' authors on the platform. In blog search, Huang et al. (2018) interpret *authoritative evidence* as the relatedness of a blogger/feed's content to controversial topics and used it as a proxy to estimate *opinion*, as we will describe later in our analysis. Controversial topics refer to those that may cause controversy, argument, and polarized opinions. Concluding, these relevance factors refer to a document's source/publisher or its content.

Coverage, specificity. Both of these concepts are related to textual content. Specifically, da Costa Pereira et al. (2009, 2012) define *coverage* as a measure related to the degree a user's interests are included in a document. A similar definition is provided by Dumitrescu and Santini (2021), who perceive it as the proportion of a user's interests represented by the documents retrieved from the stream, that is, news streams, within a specific time span. Singh and Dave (2019) characterize minimum coverage as the shortest segment of the document, which covers all the user query terms that appear in that document. In math search, Zhang and Youssef (2014) estimate coverage by measuring the portion of a mathematical expression mentioned in a query and a given document. Shajalal and Aono (2020) describe coverage in the context of their study as a measure that considers both the relevance of a subtopic to a query and how

frequently that subtopic appears in documents. *Specificity*, in the context of XML retrieval, refers to how focused an XML element is on the topic of request, meaning it does not discuss other topics, irrelevant to the user's query (Ashoori & Lalmas, 2007). According to some scholars these relevance factors describe either a user-document or a query-document relationship, although they are refereed to with the same terminology.

Spatial relevance, location, distance. Each of the terms highlighted above relates to geographic locations; however, our analysis will explore their specific interpretations. Specifically, Daoud et al. (2013) assess *spatial relevance* by concentrating on the query intent rather than the actual geographic *location* of the user, which is the focus of studies by Kang et al. (2012); Bouidghaghen et al. (2011); Moulahi, Tamine, and Yahia (2014). Therefore, these relevance factors can be query- or user-dependent.

Objectivity, opinionatedness, opinion. In the context of web search, Lioma et al. (2016) use the notion of *objectivity* along with the concept of *factuality* as proxies to credibility. The authors consider objectivity as the degree to which the text meaning depends on the author's perspective, that is, the exact opposite notion of subjectivity. Regarding the concept of *opinionatedness*, Putri et al. (2020) define it based on the likelihood of a document to express an opinion about a query, a synonym of the term *opinion*. However, Huang et al. (2018) assume that an *opinion* score reflects the extend to which a blog post is about controversial topics. Therefore, these relevance factors describe a relationship between a query and a document, or can be solely a document's attribute.

Content quality, passage quality, web page quality. According to Bendersky et al. (2011), *quality* of a web page can be evaluated based on multiple factors including its originality, trustworthiness, content relevance, metadata accuracy, interlinked resources, and user-centric layout design. From the provided description, it is evident that the concept of quality is broad, incorporating multiple of the previously described relevance factors. The domain of community question answering has also utilized the concept of *passage quality* regarding the retrieved answers (Amancio et al., 2021; Yulianti et al., 2018). Nonetheless, the domain has yet to offer a clear definition of the concept of quality. While assimilating multiple relevance factors under the concept of quality, it remains consistent that this notion pertains solely a document's attribute.

Popularity, PageRank, reputation. From our review of the included studies, the concepts of *popularity* and *reputation* emerge within e-commerce, social search, web search, and local search contexts. Badache and Boughanem (2014) treat them as two distinct notions that characterize a document, and define popularity as a measure of how well-known a resource is among the public, primarily driven by sharing and commenting activities on social networks; while reputation reflects the general opinion or appreciation of that resource, determined by positive social actions, such as number of likes. In e-commerce, Bassani and Pasi (2021) exploit products' popularity as ranking feature. A product's popularity is reflected by how often users choose it. *PageRank* has been defined by Brin and Page (1998) as a measure that quantifies the importance or popularity of a web page based on the number and quality of links pointing to it. Based on their definitions, these relevance factors are document attributes whose estimation relies on user actions on documents or reflects general appreciation based on huge amount of users.

Even though we have made significant efforts to combine all the relevance factors mentioned in Table 1 based on their conceptual similarity, there remain certain domain-specific factors, like document's usage and citations in legal search (Wiggers et al., 2023), which we could not assimilate with other factors. Therefore, we direct readers interested in these specific factors to the original papers.

4.2.2 | What methodologies are used to operationalize the identified relevance factors?

Based on Table 2, we address this research question by presenting and discussing the methodologies that have been leveraged to operationalize the identified relevance factors (i.e., estimate a related score).

Topicality. In the majority of the included studies, *topicality* has been estimated using lexicon-based retrieval models, commonly the BM25 model (Robertson et al., 1994). Studies employing a LtR approach utilize multiple lexicon-based retrieval models to estimate topicality scores, subsequently incorporating them as input features within the LtR models. Among the reviewed studies, the LambdaMART algorithm, developed by Burges (2010), emerged as the most frequently employed LtR approach.

Appropriateness, user's interest, user-based relevance, appropriateness for children, user intent, user's habit, user's familiarity. *Appropriateness* has been calculated based on the similarity of term-based vector representations of a given document and user's interest (da Costa Pereira et al., 2009, 2012). A user's interest is based on user-related information such as a set of authored documents or a personal description. So, the notion of appropriateness

encompasses the user's interest. Bouidghaghen et al. (2011) estimate a *user's interest* on a document as the sum of cosine similarity scores between the document's vector representation and vectors representing k concepts extracted from a user profile. These concepts have been selected based on the Open Directory Project.[1] Tamine et al. (2011) address the challenge of capturing a user's interest in academic search. They employ social network analysis to construct a graph connecting authors and their publications. Leveraging the assumption that co-authors share similar interests, they compute scores based on the graph, reflecting co-authorship, citations, and authorship. The document's interest score is obtained as a weighted sum of the three scores for each author-document pair. Li et al. (2017) and Uprety et al. (2018) investigate features related to user interest. They calculate three cosine similarity scores between a document's vector representation and a concatenated vector representation of all SAT-clicked documents within different time frames (session, day, and long-term). In addition, they extract topics from SAT-clicked documents using Latent Dirichlet Allocation. Then, they construct concatenated document representations in each time frame whose elements indicate the probability that the document is relevant to specific topics. Finally, they estimate three cosine similarity scores for each document based on its topic-based vector representation and the concatenated SAT-clicked document representations. Sieg et al. (2007) assume that the notion of user's interest changes during a search. They create an ontological user profile, updated during the search session to reflect changes in the user's interests. Users' interests are concepts extracted from the Open Directory Project that are updated based on the user's behavior during search (e.g., visited web pages, time spent on a web page). Also Sahraoui and Faiz (2017) consider the user's interest as a dynamic notion during a search session. The authors estimate users' interests implicitly from their social web activities and represent them as vectors of weighted terms. Recognizing the evolving nature of interests, they adjust term weights based on the recency and frequency of the web activities they occur. By doing so, the approach captures new and persistent interests. Dumitrescu and Santini (2021) introduce a set of algorithms to dynamically filter a stream of documents, ensuring they align with a user's interests and provide a diverse range of content. To achieve that, they leverage the Self-Organizing Map algorithm, creating a user model from a representative collection of documents. Their approach distinguishes new content from areas users have not recently engaged with and ensures comprehensive coverage to their varied interests. Moving to the *user's habit* factor, it has been operationalized by Li et al. (2017) and Uprety et al. (2018) using three different methods that leverage behavioral signals extracted from query logs. The first evaluates users' preference for a particular source website, drawing from their historical interactions. The other models capture user's preference toward documents based on specific lengths and language. In e-commerce, Mandayam Comar and Sengamedu (2017) operationalize *user intents* and incorporate them in their relevance model by looking at how often users clicked on results at different positions, estimating the click-through rate (CTR) of user profiles. Users with purchase intent typically have a rapidly declining CTR as position increases. In contrast, exploration intent shows a consistent CTR across positions. Summarizing the aforementioned studies one can observe two distinct computational methodologies. The first treats these relevance factors as static (per user) and mainly relies on similarity measures to estimate a document's interest to a user profile. These approaches leverage user and document contents or a user's social network. The other category of approaches assumes that a user's interest changes during search and over time. These methods estimate interest based on users' actions on documents (or products), for example, dwell time, visiting history, among others. While these approaches have not been experimentally compared to each other, the second approach, which encompasses the first, appears more promising due to its adaptability to evolving user interests and real-time behavior.

Reliability, credibility, trustworthiness, genuineness, factuality and objectivity, correctness. A source's *reliability*, according to da Costa Pereira et al. (2009, 2012), could be estimated based on the degree that a user trusts a document's source. To estimate reliability, Li et al. (2017) and Uprety et al. (2018) employ seven features based on SAT-clicks. The first feature relies on the number of SAT-clicks a document receives; the more clicks it accumulates, the greater its perceived reliability. Similarly, the second feature estimates the number of SAT-clicks on the document's source. The third feature estimates the ratio of clicks and SAT clicks on a document, and the fourth is the same ratio based on a document's source. Three additional scores are obtained by well-known methods on the literature that correspond to a document's PageRank, predicted reliability score, and spam score. Notably, the authors leverage the PageRank score of a web page as a proxy for its reliability. Fernández-Pichel et al. (2022) argue that a document's reliability needs to be estimated primarily relying on query-related document's content. To estimate a reliability score, the authors propose two approaches, one based on a fine-tuned Mono T5 model that classifies a passage as reliable or unreliable and an unsupervised approach that measures the similarity of a document's passage to true and false query-related handcrafted claims. In academic search, Jomsri and Prangchumpol (2015) associate a document's reliability based on the type of research paper publication, which varies from published in a journal to uploaded as a file on the web. In social search, Putri et al. (2020) exploit a model-driven approach based on MCDM, initially proposed by Pasi

and Viviani (2018), to estimate a document's credibility score. For each tweet, the authors estimate features such as the number of followers, the number of URLs, the number of retweets, and the author's account age. The obtained scores are aggregated into a single credibility score using ordered weighted averaging operators. To estimate a *trustworthiness* score, Ravikumar et al. (2013) use a set of features related to a user's profile (number of followers, verified profile, among others) and to the content of the tweet, for example, length, or hashtags. All these features are used in a LtR setting to predict an overall score. To estimate credibility/trustworthiness in consumer health search, Putri et al. (2021) leverage a set of features related to the presence (i.e., the number) of internal and commercial links in a document. Additional features are related to the page's presentation, existence of advertisements, and the PageRank score. We refer the reader to the original publication for a complete list of the exploited features. In the same retrieval task, Upadhyay et al. (2023) introduce a BioBERT model fine-tuned based on labels that account for both topicality and credibility (i.e., the model estimates relevance based on both factors). The model is used to re-rank a set of documents initially retrieved based on their topical relevance using the BM25 model. Banerjee et al. (2023) leverage credibility, correctness, and topicality for online health IR. In their approach, the authors leverage summarized document representations. A document's correctness score is calculated based on the maximum cosine similarity between its embedding representation and the embedding representation of a query expression derived from a combination of its description and its answer (provided in the leveraged dataset). To estimate a credibility score, they train a logistic regression model that predicts a binary score for each document. Upadhyay et al. (2022) propose an unsupervised method to evaluate the genuineness of online health information using a set of scientific articles that can support the claims made in a document. The authors compute a genuineness score by estimating and aggregating the cosine similarity values between the context of a document and a set of published scientific medical articles that cover the same topic.

Finally, Lioma et al. (2016) estimate credibility based on indicators of *factuality* and *objectivity*. The authors constructed two distinct data collections and trained two support vector machine (SVM) classifiers that predict these scores. The estimation of these factors relies mainly on the document's source. The most notable observation is the usage of PageRank as a proxy to estimate a web page's reliability. At the same time, some scholars estimate it from a user's viewpoint (da Costa Pereira et al., 2009, 2012; Putri et al., 2020). Finally, one can notice that the estimation of credibility and trustworthiness in social search leverages similar features, yet the terminology is different.

Freshness, temporal relevance, recency, novelty. To estimate multidimensional relevance incorporating the notion of *freshness*, Badache and Boughanem (2014) propose a model that relies on counting specific social actions (i.e., like, share, comment) conducted on a resource (i.e., document). This model adjusts the count based on when an action occurred, so resources with more recent actions are ranked higher. Based on their domain-specific definition Bambia and Faiz (2015) assume that *freshness* can be described by a set of known terms extracted from current search trends or other sources. Then, using a language model, the authors evaluate the closeness of query terms to those terms and estimate a freshness score for each document. In their study, Dai et al. (2011) assess freshness by creating features that leverage a temporal contextual profile of queries. These features are constructed based on a set of documents retrieved for a given query based on topical relevance. Such a feature is a document's temporal PageRank score; for a complete list of the exploited features please refer to the original publication. Dumitrescu and Santini (2021) exploit the notion of freshness alongside the notions of user's interest (aka personalization) and coverage. To incorporate the notion of freshness in search, the authors integrate the timestamp of an item in their estimations. Similarly, in academic search, Jomsri and Prangchumpol (2015) integrate the *recentness* of a publication into their ranking, utilizing a normalized version of the paper's publication year. In social search, Moulahi, Moulahi, et al. (2014) estimate a tweet's recency by considering the time lapse between its publication and the submission time of a query. Likewise, in community question answering, Amancio et al. (2021) employ features like the answer's creation date, the most recent date mentioned in a referenced web page text, to train a LtR model to associate a recency score for an answer. Jabeur et al. (2012), although they do not define the notion of *temporal relevance*, they estimate it based on the occurrence of query terms in temporal neighbor tweets under predefined temporal intervals. Yang et al. (2021) propose a domain-specific notion of temporal relevance, namely *seasonality* of products. Even without a formal definition, this notion is intuitively understood. To predict seasonality, the authors train a LtR model that utilizes the annual sales data for a calendar year and create vector representations based on product–month relationships. To estimate a temporal relevance score in geographic IR, Daoud et al. (2013) use a probabilistic ranking model that considers the temporal frequency of terms within the document and the weight of the temporal query context. To estimate *novelty* in their models, Li et al. (2017) and Uprety et al. (2018) exploit four features grounded in both temporal and psychological views of novelty. As a result, the authors estimate novelty with respect to both the user-document and the query-document temporal relationships.

User-based novelty is calculated based on the divergence between the language model of a retrieved document and those of previously viewed documents using the Kullback–Leibler divergence. Another feature adds a forgetting factor to the previously mentioned feature. The third feature relies on the number of words that appear in the retrieved documents and a user's SAT-Clicked documents. Finally, the last feature relies on the difference between a document's production and retrieval times. Based on the reviewed studies, one can observe that the majority leverages a document's metadata (e.g., timestamp, publication year) to estimate these notions. Other scholars leverage document and query contexts, while others estimate them based on user actions (i.e., SAT-Clicks). Although the proposed approaches differ and are not directly comparable, the features exploited by Li et al. (2017) offer the most holistic estimation.

Readability, understandability. These relevance factors have been employed in web search (Collins-Thompson et al., 2011; Li et al., 2017; Sasaki et al., 2016; Uprety et al., 2018), and in the health (Palotti et al., 2019; Putri et al., 2021; van Doorn et al., 2016; Zhang et al., 2015), and the academic (Arastoopoor, 2018) domains. In web search, Collins-Thompson et al. (2011) employ three approaches to estimate a document's *readability* score based on its snippet (as appeared in a search result page) and body texts, and one approach that estimates readability based on a user's proficiency level. These, along with features related to topicality, are utilized for training a LambdaMART model for web page retrieval. Initially, a language model assigns a readability score to a document based on the percentage of words familiar to a percentage of the general population. The second approach involves calculating each web page's Dale–Chall reading difficulty measure (Chall & Dale, 1995). The last approach employs a Logistic regression classifier to label each web page, signifying its topic (e.g., Kids & Teens category). Finally, users' reading proficiency is inferred from their search behavior using a probabilistic framework and leveraging clicked documents in a session and a user's frequently visited web pages. Sasaki et al. (2016) exploit a logistic regression model to estimate a document's readability. The input features used for training are encompass the average number of syllables per sentence, the Dale–Chall measure, the rates of unigram and bigram-based POS tags, the average sentence length, the depth of heading tags in the web page, and the document length. Both Li et al. (2017) and Uprety et al. (2018) leverage the same content-based features to estimate a web page's understandability, incorporating these features directly into their proposed retrieval models. These features include the count of easy and difficult words in a document, the reciprocal average word length, as well as the well-known readability formulas, namely the SMOG Index, Coleman–Liau Index, Gunning Fog Index, and the Flesch–Kincaid. In the health domain, Zhang et al. (2015) introduce a two-step approach to estimate a document's readability based on its content and underlying topics. Initially, their methodology extracts all topics from a document collection using the hierarchical Latent Dirichlet Allocation approach. Subsequently, the authors estimate three scores: a Topic Trace score, calculated by tracking a document's topics sequentially on the taxonomy; a Topic Scope score that reflects the coverage of the identified topics in a document; and a Dale–Chall score. These scores are aggregated into a single readability score based on a custom formula proposed by the authors. To estimate understandability, van Doorn et al. (2016) train an SVM model using a document's Coleman-Liau index, Gunning fog index, the number of medical term occurrences, and a document's length as input features. Palotti et al. (2019) conduct a comprehensive investigation into methods for estimating the understandability of health-related web pages. Their findings indicate that machine learning techniques outperform traditional readability metrics, specifically the XGB regressor, which leverages natural language, HTML structure, and domain-specific features. Interested readers are referred to the original publication for a complete list of the features used as inputs. Nonetheless, the use of well-known readability formulas (i.e., those mentioned above) to estimate understandability in the medical domain continues to be a common practice (Putri et al., 2021). In the academic domain, Arastoopoor (2018) explores the use of classic readability measures in Persian scientific texts. Each document is assigned a readability score using a modified version of Flesch–Dayani's formula tailored to the Persian language. The fundamental definitions of readability and understandability emphasize their reliance on user perception. However, among the reviewed studies, only one approach directly considers the user's understandability level. In contrast, others estimate it based on general population metrics using readability formulas. Subsequently, most studies treat understandability estimation as a document property rather than a user-centered relevance factor. That underscores a potential limitation in estimating a document's understandability, as it may overlook nuanced factors that contribute to users' comprehension and accessibility of documents' content. Finally, despite the cost-effectiveness and simplicity of well-known readability formulas, machine learning regression models lead to superior performance in estimating understandability, particularly in the health domain.

Content diversity, exhaustivity, scope. To achieve *diversity*, van Doorn et al. (2016) employ a cluster-based ranking technique that uses a bag-of-word representation for documents. This technique re-ranks a set of documents, which were initially ranked based on topical relevance according to their topical diversity. Xu et al. (2016) use a group-wise LtR framework that retrieves topically relevant and diverse documents. Their model relies solely on features related to

topically, while diversity has been incorporated during the training phase. Specifically, during the training phase, the authors divided relevant documents into groups based on the different aspects they covered. Each group consisted of a document that covered more aspects (labeled as 1) and several others with fewer aspects (labeled as 0). The document with more aspects encompassed all the aspects found in the other, less comprehensive documents. Vargas et al. (2012) propose a probabilistic relevance model that unifies the IA-Select and xQuAD models aiming to integrate result diversification based on users' intent. To estimate *exhaustivity*, Ashoori and Lalmas (2007) leverage a topic segmentation algorithm based on lexical cohesion (i.e., the one employed to measure specificity). This algorithm captures the number of topics in an element by measuring the changes in the employed vocabulary. Even though Li et al. (2017) and Uprety et al. (2018) leverage the *scope* relevance factor that encompasses both coverage/specificity and exhaustivity, they operationalize solely its coverage aspect. The authors construct a set of features for their models. The first feature is the Jaccard Index between query and document topics. The second feature measures the number of document passages that contain query terms. The last feature is the sum of the cosine similarities of the embedding representation of each document word to a concatenated embedding representation of a query. The aforementioned approaches employ a diverse range of strategies to achieve content diversification within a retrieved set of documents, with most focusing on leveraging both query and document content.

Authority, authoritative evidence. Farah and Vanderpooten (2008) estimate a web page's *authority* based on the number of incoming "good" links to the web page. The studies by Zhuang et al. (2021) and Eickhoff and de Vries (2014) also mention the notion of authority. However, this notion is operationalized in their systems by a web page's characteristics, such as its PageRank score, quality, among others. The score is not computed; instead, it serves as one of the document features within the dataset (i.e., Microsoft's WEB30k) utilized for experimental model evaluation by the authors. Moulahi, Moulahi, et al. (2014) estimate a user's authority on Twitter by summing the number of tweets the user has published and the number of times the user has been mentioned or cited by others. Therefore, the notion of authority is estimated using the web page's PageRank score in web search, while in social search, it is determined by users' attributes.

Coverage, specificity. In newswire search, da Costa Pereira et al. (2009, 2012) estimate coverage by creating vector representations of words appearing in documents and user profiles. A coverage score is computed using fuzzy inclusion, which relies on the cardinalities of the fuzzy subsets representing a user's interests and a document's content. For Dumitrescu and Santini (2021), coverage is a notion estimated for a set of documents based on a user profile and using a standard Self-organizing Map algorithm. In their approach, as content related to a topic is engaged during search, the interest in that topic and nearby semantic areas, that is, the other documents in a retrieved set, decreases. After a series of items is examined, the overall level of remaining interest indicates the coverage of those items. In academic search, Singh and Dave (2019) leverage query and document contents to estimate a coverage score. They propose a formula that divides query length by the product of the number of query terms missing from the document and the length of the shortest document segment that covers all query terms. Mathematical coverage has been estimated by counting the number of query terms (i.e., mathematical expressions) covered in a document (i.e., mathematical formula) (Zhang & Youssef, 2014). In web search, Shajalal and Aono (2020) introduce a formula that estimates coverage by multiplying two factors. The fraction of visible text terms (rendered in a web browser) to the total number of document's terms (i.e., information-to-noise ratio). The document's ranking position in search results, estimated by BM25 and normalized by the total number of documents. In this coverage estimation approach, one can notice that the notion of topicality is encompassed in that of coverage. To measure *specificity*, Ashoori and Lalmas (2007) utilize a topic segmentation algorithm based on lexical cohesion. This algorithm captures the number of topics in a document by measuring the changes in the employed vocabulary. Based on these studies, the notion of coverage is estimated through user-document or query-document relationships relying on similarity measures on vector representations or term occurrences. On the other hand, specificity is a document's property estimated by the number of topics it covers.

Spatial relevance, location, distance. In geographic IR, to estimate the *spatial relevance* of a document to a query, Daoud et al. (2013) first extract the query's geographic context (i.e., locations) from a set of pseudo-relevant documents retrieved based on topicality. The geographic score of a document is determined using a probabilistic ranking model, where instead of inverse document frequency, the frequency of documents with a geographic expression is used. To estimate spatial relevance Palacio et al. (2010) create a domain-specific index architecture in which each document is represented by its spatial information, that is, location-related entities. Using this index, the authors compute a relevance score (similarity score) reflecting the spatial relationship between the documents and the query. Kang et al. (2012) consider three relevance dimensions, namely topicality, *location*, and reputation in the context of local search. The authors propose two methodologies; the first trains a LtR model using label aggregation across the three relevance

aspects so that the model retrieves documents that satisfy all the considered dimensions simultaneously. The second approach predicts a score associated with the distance relevance dimension based on a linear regression model. In the context of mobile search, Bouidghaghen et al. (2011) use a geographic weighting function introduced by Wang et al. (2005). Given a user's location (e.g., a city), a document, and a geographic hierarchy, a relevance score is calculated by adding the occurrences of the user's location and the offspring locations in the document. Moulahi, Tamine, and Yahia (2014) incorporate the concept of *distance* in contextual search aiming to retrieve a set of places (e.g., restaurants) for a given user. The proposed approach computes a score based on the actual geographic distance between the user's location and the location of each place. In conclusion, each study employed a unique method to estimate these relevance factors. However, a shared assumption is that the location of interest is consistently included in the query text and, depending on the context, it might not represent the actual physical location of the user.

Objectivity, opinionatedness, opinion. Lioma et al. (2016) investigate how the notion of document's *objectivity* impacts retrieval effectiveness in newswire and web searches. Relying on the approach proposed by Wiebe and Riloff (2011), the authors leverage a set of predefined subjective or objective nouns, a lexicon, sentence syntactic patterns, and part-of-speech features, to train a sentence objectivity SVM classifier. The overall document's objectivity is computed by considering the ratio of objective sentences to the total number of sentences within the document. In social search, Putri et al. (2020) estimate *opinionatedness* following the approach proposed by Giachanou et al. (2016) that linearly combines a term-based and a stylistic-based opinion scores. The term-based score is derived from the average opinion score across all terms within the document, utilizing the AFINN Lexicon to identify opinionated terms. The stylistic-based score is calculated based on the frequency of emoticons, exclamation marks, character repetition, and hashtags in a tweet and in the collection. In blog post search, Gerani et al. (2012) estimate a document's *opinion* leveraging the proximity-based opinion method introduced by Gerani et al. (2010). This method initially calculates the probability of opinion in different passages of a document and then estimates an overall probability leveraging a proximity-based density kernel. Eickhoff et al. (2013) compute documents' (blog posts') opinion scores using the LingPipe classifier. However, the original paper lacks further details on this specific aspect of their methodology. In their work, Huang et al. (2018) estimate an opinion score for blog posts based on their association with controversial topics. Using a training set of blogs, the authors identify terms that express controversial topics in the collection and assign them an opinion weight based on the Kullback–Leibler divergence formula. Then, an opinion score to each blog post is given by a language model, and specifically by summing the logarithm of the generation probability of the top weighted topical terms. These relevance dimensions have been estimated through various computational approaches, ranging from trained classifiers to simpler score aggregation of related features. Although the majority of reviewed studies rely on lexicons and textual features, Huang et al. (2018) take a different approach by leveraging terms associated with controversial topics, showcasing a diverse methodology in their relevance estimation.

Content quality, passage quality, web page quality. In web search, Bendersky et al. (2011) incorporate a document's *quality* into their retrieval method by utilizing a comprehensive set of features that encompassed aspects related to a web page's content, structure, and presentation within a web browser. An overall quality score is computed through a weighted sum of several features, including the total number of visible terms within the body section of a web page, the total number of terms in the title, the average length of all visible terms (that is an indication of content's readability), the fraction of anchor text relative to all of the web page's text, the fraction of anchor text over all visible terms, the ratio of stop words to non-stop words, the fraction of stop words in the visible text, the fraction of table text on the web page, the depth of the URL, and the computation of an entropy score derived from all of the web page's terms. The computation complexity of calculating these features is linear to the number of a document's terms. Eickhoff and de Vries (2014) and Zhuang et al. (2021) propose retrieval models for web search that incorporate a web page's quality. For their experimental evaluation, these studies utilize Microsoft's WEB30k dataset, where each web page is assigned two quality scores derived from two distinct classifiers. Due to the absence of comprehensive details in the original papers, providing further descriptions of these classifiers is not feasible. In their research on community question answering, Yulianti et al. (2018) incorporate *passage quality* in the retrieval process. The overall passage quality score is calculated based on the weighted sum of the following features, some of which have also been used by Bendersky et al. (2011). These features encompass the number of sentences in a passage, the number of query terms present in a passage, the average passage term weight and term length, a passage entropy score, the fraction of stop words in a passage, the fraction of passage terms that are stop words, the score assigned to the best matching passage in the retrieved document, and the number of overlapping bigrams between a passage and its related answers. Also, for community question answering, Amancio et al. (2021) introduce an LtR model that leverages 186 features to estimate the quality of each document (i.e., answer). These features are classified into textual that are related to the structure of answers (e.g., number of

sections, code snippets, and images), text size (such as word count), writing style, and correctness (encompassing readability-related features). Moreover, the quality prediction model incorporates nontextual features, such as the reputation of the user who submitted the answer or the number of edits an answer has undergone. For a comprehensive list of quality-related features, refer to the original publication. In summary, quantifying a document's quality has been approached through LtR methods or classifiers, which utilize a combination of textual or nontextual features or by estimating the weighted sum of a smaller set of textual features. We also observe that a document's quality is closely entangled with a text's readability.

Popularity, PageRank, reputation. In e-commerce, *popularity*, as integrated by Bassani and Pasi (2021) in their retrieval approach, is derived from an item's total number of purchases (i.e., it is a product's attribute). To mitigate bias toward popular products, they use the *n*-root of the total purchases. In social search, Badache and Boughanem (2014) combine popularity and reputation dimensions into their methodology for movie retrieval. Popularity is determined by evaluating the number of comments, tweets, and shares a movie has on social platforms like Facebook or Twitter. The PageRank score is originally incorporated in Google's ranking algorithm Brin and Page (1998) combined with topicality. Craswell et al. (2005) investigated which score transformation can be applied on PageRank and lead to better retrieval effectiveness when combined with topicality scores (i.e., BM25). *Reputation* is estimated through the number of likes, mentions, and bookmarks the movie accumulates on such platforms. In their first proposed retrieval approach, Kang et al. (2012) introduce a LtR approach relying on label aggregation (as analyzed before). Their second LtR approach incorporates a document's reputation score using a linear regression model trained based on a review's rating score and number of ratings. These studies reveal that the considered relevance dimensions are operationalized through users' interactions with documents. Depending on the application domain, these interactions are measured using inherent document attributes such as the number of likes, shares, reviews, or purchases.

The relevance factors previously mentioned encompass most of the factors that have been identified and implemented to estimate multidimensional relevance in the reviewed studies. Despite the fact that we make a big effort to merge them based on their conceptual similarity, there are still some domain-specific factors such as *document's usage*, and *citations* in legal search (Wiggers et al., 2023), that we could not merge with other factors. For these factors, we refer the interested readers to the original publications.

4.3 | Which benchmark collections have been used to estimate multidimensional relevance, and how are they characterized based on their annotated relevance factors, availability, and size?

This section presents the benchmark collections employed to evaluate multidimensional relevance models in the reviewed studies. Each collection is described based on its source, availability, and characteristics (number of queries, documents, and relevance judgments). Further analysis is conducted based on the relevance factors investigated using each collection. Lastly, in case several studies evaluate their approaches on the same collection, we comment on their achieved performance. By presenting these aspects, we aim to offer a comprehensive understanding that can support researchers and practitioners in being aware of the available resources and advancing this domain of study.

Analyzing the 72 reviewed studies, we identified 41 studies that use benchmark collections constructed within initiatives such as TREC,[2] Conference and Labs of the Evaluation Forum (CLEF),[3] and NII Testbeds and Community for Information Access Research (NTCIR)[4] for evaluation. Three studies evaluate on collections that are not publicly available, while six use collections that are unrelated to the aforementioned evaluation campaigns. In 22 studies, the authors constructed their own collections, that is, custom collections, on which they evaluated their models. In the remaining section, we analyze the collections associated with initiatives categorized by their knowledge domain as there are publicly available and used in various of the reviewed studies.

In web search, most of the collections originated in the TREC Web Track, which was running between 1999–2004 and 2009–2014. Craswell et al. (2005) evaluate their model using the collection created in the TREC Web Track 2004. The collection relies on the .GOV dataset and contains 225 queries with 88,000 relevance judgments based on topicality. The collections created for the *TREC diversity task* in 2009, 2010, and 2012 have been used to evaluate models that exploit topicality and diversity factors (van Doorn et al., 2016; Vargas et al., 2012). These collections leverage the multilingual ClueWeb09 dataset that contains about 1 billion web pages collected in January and February 2009. Each collection provides 50 queries and has 28,000, 9000, and 62,000 relevance judgments based on diversity and topicality. The collections created for the *TREC* ad hoc *web tracks in* 2009, 2010, 2011, and 2012 have been used to evaluate

multidimensional models that exploit topicality, document quality, readability, factuality, and objectivity factors (Bendersky et al., 2011; Lioma et al., 2016; Sasaki et al., 2016; Yulianti et al., 2018). These collections use the ClueWeb09 dataset, and when combined, they provide 200 queries (50 each) with almost 83,000 topicality-based relevance judgments. Lioma et al. (2016) evaluate their model on the TREC ad hoc test collection that consists of 150 queries and has 144,000 relevance judgments based on topical relevance. Yulianti et al. (2018) use their quality-based relevance model on the collections created in the TREC Terabyte tracks 2004–2006. These collections leverage the. GOV2 dataset and contain 150 queries with almost 135,000 relevance judgments. The collection provided in the topic distillation task in TREC Web Track 2013 have been used for evaluation by Farah and Vanderpooten (2008). The collection leverages the clueweb12 dataset, comprising 733 million English web pages crawled in 2012; it encompasses 50 queries and 14,000 topicality-based relevance judgments. Uprety et al. (2018) and Li et al. (2017) evaluate their models leveraging the collections created in the TREC session tracks 2013 and 2014. These collections use the clueweb12 dataset and provide 69 and 60 queries with 13,000 and 16,000 topicality-based relevance judgments, respectively. The proposed models estimate relevance based on seven factors, such as novelty and understandability. However, relevance assessments across the collections rely solely on topicality. Regarding the achieved retrieval performance, both models attained identical levels of performance. Zhuang et al. (2021) evaluate their approach on the Yahoo! Learning to Rank Challenge data (Chapelle & Chang, 2011). The dataset contains around 882,000 query-document pairs represented as vectors of features and relevance judgments. These vectors comprise a wide range of features associated with topicality, authority, and popularity. Similarly, the MSLR-WEB10K and the MSLR-WEB30K have been used to evaluate multidimensional relevance models that estimate relevance based on topicality, web page authority, quality, among others (Chapelle & Chang, 2011; Eickhoff & de Vries, 2014). Shajalal and Aono (2020) evaluate their model using the NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 collections that comprise 130 and 67 million Chinese and Japanese web pages and contain queries in English, Chinese and Japanese. We observe that relevance judgments across the collections employed in web search are mainly based on topicality, except the one introduced in TREC's diversity task. However, relying solely on topical similarity for judgments may only capture a fraction of the complexity of relevance perceived by users in real-world scenarios. As a result, the retrieval effectiveness of multidimensional models that consider user- and task-related factors might be underestimated compared to systems that rely their relevance estimation solely on topicality.

In the medical domain, numerous studies introduce multidimensional relevance models that consider topicality, readability, understandability, credibility, and trustworthiness factors (Banerjee et al., 2023; Palotti et al., 2019; Putri et al., 2021; Upadhyay et al., 2023; van Doorn et al., 2016; Zhang et al., 2015). These models have been evaluated on datasets introduced in the CLEF eHealth retrieval tasks 2013, 2015, 2016, 2018, and 2020. The 2013 and 2015 collections comprise the same 1 million health-related web pages and have 50 and 67 queries, respectively. Regarding their relevance judgments, the 2015 collection contains 12,000 judgments based on readability and 8000 based on topicality. The number of relevance judgments in the 2013 collection is not explicitly mentioned. The 2016 collection leverages the clueweb12 collection and contains 150 queries and 25,000 relevance judgments based on topicality, understandability, and trustworthiness. Finally, the 2018 and 2020 collection contains 5 million medical web pages and each contains 50 queries. Although the number of relevance judgments in the 2020 collection is not explicitly mentioned, the 2018 has 26,000 relevance judgments based on topicality, readability, and trustworthiness. Among the studies utilizing these collections, only the works by van Doorn et al. (2016) and Palotti et al. (2019) evaluate their approaches using the same dataset. According to reported retrieval effectiveness, the result fusion method proposed by Palotti et al. (2019) shows superior retrieval performance in terms of the understandability rank-biased precision measure. The collections introduced in TREC Precision Medicine tracks in 2017, 2018, and 2019 have also been used to evaluate multidimensional relevance models. These collections have two document types, namely scientific abstracts and clinical trials and include 30, 50, and 40 queries, respectively. Regarding the available relevance judgments, they have around 22,000, 21,000, and 18,000 relevance judgments based on topicality. Xu et al. (2016) evaluate their model on the collections introduced in TREC Genomics tracks 2006 and 2007 that contain HTML documents from medical journals. In addition, these collections provide 28 and 36 queries with 27,000 and 35,000 relevance judgments, respectively. The authors also use TREC Clinical Decision Support tracks 2014, 2015, and 2016 for evaluation. These collections use medical case narratives to retrieve biomedical articles and contain a total of 90 queries and almost 114,000 relevance judgments. In both tracks, relevance has been assessed based on topicality. The TREC Misinformation track 2020, focusing on COVID-19 misinformation, aims to retrieve useful, credible, and correct information. The document collection consists of CommonCrawl news articles sampled from January 1, 2020 to April 30, 2020, specifically focusing on health-related news worldwide. There are 50 queries associated with this collection, accompanied by 21,000 relevance judgments based on topicality, credibility, and correctness. Finally, CLIREC is a test collection for evaluating clinical IR that exploits a set of

manually crafted PICO-structured queries to retrieve medical documents (Znaidi et al., 2016). The collection comprises 1.2 million documents gathered from PubMed, 423 queries and has 8000 relevance judgments. In contrast to collections utilized in web search, we observe that a higher proportion of collections contain relevance judgments associated to other relevance factors.

In social search, we have identified three benchmark collections. The first is introduced in the CLEF Microblog Cultural Contextualization Lab (Ermakova et al., 2017) and contains 70 million event-related micro-blogs collected over 18 months and 53 queries. The other two collections comprising 16 million tweets are introduced in TREC Microblog tracks 2011 and 2012. They contain 60 queries each, 73,000 and 60,000 topicality-based relevance judgments, respectively. In e-commerce, Bassani and Pasi (2021) leverage the Amazon review five-core dataset to evaluate their model that relies on topicality, popularity, and other task-related factors. The dataset comprises thousands of queries, millions of products (i.e., documents), and users. In blog post search, several studies leverage the TREC Blog tracks (2006–2010) to evaluate their models (Chenlo et al., 2015; Eickhoff et al., 2013; Gerani et al., 2012; Huang et al., 2018). In newswire search, the Reuteurs RCV1 Collection has been employed to evaluate models that rely on topicality, coverage, appropriateness, among others factors (da Costa Pereira et al., 2012, 2009; Dumitrescu & Santini, 2021). Each publication within the remaining knowledge domains listed in Table 1 utilizes a single collection to assess the proposed model. Furthermore, because of the absence of suitable benchmark collections, numerous studies in academic, legal, educational, expert finding, local, mathematical, and mobile searches rely on custom collections for evaluation. This circumstance complicates research in the field since it is challenging to develop and compare retrieval models effectively. Nonetheless, we encourage interested readers to consult the original publications for additional details regarding these collections.

Based on the analysis of the reviewed studies and the benchmark collections they employ for evaluation, several observations have emerged. The most prominent observation is the need for available datasets with annotation based on several relevance factors. This need becomes more evident as we observed a pronounced correlation between the amount of research studies in a particular domain and the availability of benchmark collections for that domain. The second observation concerns the collections presently employed for evaluation purposes. The proposed multidimensional relevance models are usually assessed solely on labels that evaluate topical relevance. Despite this evaluation approach, the reviewed multidimensional models enhance system performance across various search tasks compared to topicality-based models. Nonetheless, having relevant judgments based on multiple factors might further highlight their potential. It is acknowledged that generating such judgments demands more time and resources. Nevertheless, the potential to develop multidimensional retrieval systems could make it worthwhile for real-world applications.

5 | DISCUSSION AND SUGGESTIONS FOR FUTURE RESEARCH

This section discusses the findings from our thorough literature examination concerning estimating multidimensional relevance. The aim is to synthesize the primary findings and underscore the contributions of this review.

Our analysis revealed that relevance is conceptualized and operationalized as a multidimensional notion across various knowledge domains and search tasks. Over the years, this research area has facilitated numerous international collaborations, maintaining a steady volume of publications. Moreover, the domain connects industry and academia, with some domains dominated by industrial contributions (e.g., e-commerce) and others, like the medical domain, by academia. Nonetheless, there are evident synergies between the two. Such collaborations underscore the theoretical interest and the substantial real-world applicability of multidimensional relevance search systems. Although our review included several diverse domains and tasks, we distinguished shared practices regarding the exploited relevance factors and the models employed to estimate multidimensional relevance.

Relevance factors. Regarding the employed relevance factors, some have a consistent presence across diverse domains. Nevertheless, a significant inconsistency in their definitions and operationalization emerged. Specifically, there were instances where relevance factors, while conceptually similar, were articulated with varying terminology. For example, factors such as credibility, reliability, trustworthiness, genuineness, authority, objectivity, correctness, and factuality. These factors have been employed in the literature to determine, up to a certain degree, whether a user should "trust" a piece of information. However, we noticed that the relationship between them exhibits a form of dynamically changing contextual dominance, meaning that one study might consider reliability to be superior to credibility, that is, using credibility as a feature of reliability, while others do the opposite. This variability complicates the

endeavor of providing formal definitions for the diverse notions, and future research should address this issue. In addition, we noticed inconsistencies regarding the computation of several factors. For example, some studies estimated reliability with respect to a document's source, that is, leveraging its attributes or metadata. Other studies measured it by considering the document's contents. Moreover, others are based on the user's perceived trust in a source, giving it a user-specific viewpoint. Similar observations have been made for other relevance factors, such as those related to the temporality of information (e.g., recency, freshness). In this case, some studies calculate it based on the document's metadata, by considering the content, and also with respect to a user's related content. Similar observations can be drawn for many relevance factors in the literature and significantly undermine any effort for homogeneity.

Attempting to address the aforementioned issue, we put forward a structured formulation for defining relevance factors. In this formulation, authors should clearly define a relevance factor and elucidate its operationalization and relationship with other relevance factors from the literature. Specifically, the authors should mention whether the considered relevance factor has been estimated with respect to user [U] (e.g., leveraging a user profile), documents [D] (e.g., leveraging documents' metadata or attributes), task [T] (e.g., follow the relevance process of the search task similar to the work by Qu et al. (2020, 2021)), content [C] (e.g., text), or other viewpoints [O]. Following this approach, introducing a new term becomes unnecessary if its estimation relies on viewpoints already covered by another concept. We argue that introducing a new concept (i.e., new terminology) requires that the concept encompasses at least one new viewpoint. In any other case, the proposed approach just amplifies the quality of estimating a concept. For example, if a study introduces a neural method to calculate readability using the content of documents, it simply offers a more refined estimate compared to traditional readability formulas that also utilize document content. Given that a new concept has been introduced, the authors should describe its relationship with other concepts in the literature, followed by a justification. Based on definitions provided in the reviewed studies, an identified relationship is quality of information $\gg$ reliability $\gg$ credibility. This relationship implies that the notion of information quality includes both the concepts of reliability and credibility and the concept of reliability also encompasses credibility. It is important to note that reliability is defined as the degree to which users place trust in a source, while credibility is dependent on the source itself. Consequently, reliability is assessed from the perspective of users [U], and credibility is assessed based on the document's metadata [D]. As a result, the estimation of information quality takes into account two factors related to both user and document viewpoints [U, D].

Aggregation approaches. Our distinction between model-driven and data-driven approaches sufficiently allowed us to classify most of the studies in our review. Model-driven strategies are rooted in explicitly defined mathematical models. Our analysis showed that while many studies propose intricate methods to calculate a relevance factor's score, they mainly use a simple linear combination to estimate a final relevance score. While alternatives like copulas and MCDM methods have been suggested, they have yet to gain the community's attention, as most recent studies still exploit a linear combination. These approaches have a tendency to prioritize transparency and interpretability, which enhances their ease of understanding. However, this preference for transparency may come at the cost of potentially lower performance and, in some cases, increased computational complexity during inference.

Conversely, data-driven approaches harness a wide range of methods to address the challenge of aggregating information in multidimensional relevance estimation. These methods generally result in improved performance across most tasks; however, this improvement comes at the cost of reduced interpretability. Based on our analysis, label aggregation ranks among the predominant approaches for multidimensional relevance estimation with LtR methods. This method provides a straightforward approach for converting a multidimensional relevance problem into a single relevance estimation problem. LambdaMART and Coordinate Ascent have consistently stood out as top-performing methods throughout the studies we reviewed. Moreover, several researchers explore new directions, like query-dependent ranking models or models that adapt to changing user behaviors. Finally, interpretable multidimensional ranking models represent another avenue that is increasingly capturing research interest, especially in domain-specific search tasks.

Benchmark collections. Our exploration points to an emerging need for benchmark collections annotated with a variety of factors across domains. That does not necessarily entail diving into exceedingly complex relevance factors. Instead, initial research efforts can be simple, focusing on exploiting relevance signals tied to document attributes. Doing so makes it feasible to further investigate how integrating these attributes impacts retrieval performance metrics, such as citations and the quality of venues in academic search. We have noted that structured, multidimensional collections play a pivotal role in shaping the research landscape. This observation is substantiated by initiatives that have created benchmark collections, like TREC, NTCIR, and CLEF, that guide the academic community's focus toward specific topics. Conversely, the industry operates independently from these trends, often addressing unique challenges and

producing original datasets. Creating benchmark collections for multidimensional relevance might be a time-consuming and expensive task. However, the emergence of LLMs offers promising potential, primarily as tools to deliver relevance annotations (Faggioli et al., 2023).

6 | PROSPECTS OF THE STUDY AND LIMITATIONS

This section outlines our study's limitations associated with the search strategy's effectiveness, the coding scheme's reliability, and potential biases inherent to our methodology. Nonetheless, we highlight the relevance and significance of this study driven by recent technological advances. The recent advent of LLMs and their relevance labeling capabilities highlight the timely significance of our review (Faggioli et al., 2023). As discussed in Section 4.3, developing models for multidimensional relevance necessitates new benchmark collections, especially for specific domains. Although the creation of these collections is both resource-intensive and time-consuming, recent studies pave the way for leveraging LLMs for annotation tasks traditionally reserved for human annotators (Faggioli et al., 2023). Nonetheless, the efficacy of these models in performing such tasks is contingent upon the quality of the prompts. Our review, especially Section 4.2.1 detailing the several definitions associated with the identified relevance factors, might be instrumental in crafting these prompts.

Another factor underscoring the significance of our study pertains to the potential of LLMs' in learning how to estimate topical relevance scores. This advancement would allow the community to transition from developing IR models centered solely on topical relevance to multidimensional relevance models incorporating user, task, and domain characteristics into a retrieval process.

Having pointed out its potential, we now turn our attention to the limitations of our study. At the initial stage of the literature review process, the existing research landscape was ambiguous and difficult to predict. Due to this uncertainty, a more expansive exploration was followed, resulting in a broad scope for the systematic literature review. This relatively broad scope has been refined due to the inclusion/exclusion criteria we have selected and the search strategy we followed in our research. As a result, it is not feasible to claim that this review includes every article that leverages more than one relevance factor (defined in Section 2) for multidimensional relevance estimation. Nonetheless, our study offers a selection of articles that touch upon diverse knowledge domains and different search tasks to provide a comprehensive summary of research surrounding this topic.

Since a precise number of papers relevant to the studied topic is indeterminate, it is challenging to assess the extent to which the included studies cover the whole population. Despite this limitation, we have endeavored to ensure that our review captures a broad and representative spectrum of the available literature. After securing our final set of included studies and examining a substantial portion of them, we conducted targeted searches on Google Scholar. These searches were focused on specific research domains (for instance, the medical domain) and particular relevance dimensions such as credibility. We then reviewed the results from these targeted searches. This procedure was replicated across domains and relevance dimensions to verify that we had identified all essential studies for our survey. By doing that, we encountered studies found in our prior searches and were subsequently either included or excluded from our review. We considered that a good indication of coverage and proceeded with our analysis. Nonetheless, future research on specific knowledge domains, particularly those underrepresented in our review, like mobile and geographic search, could uncover additional pertinent studies.

Another limitation is related to the application of the coding schema. While the schema was straightforward to apply for specific attributes of the paper (such as publication year and affiliations), its application became more subjective for other aspects, like those related to the relevance factors. As a result, studies that provided a formal definition were more lucidly represented than those that did not. In addition, there were instances where papers did not comprehensively detail the tools and methodologies they utilized. This lack of full disclosure posed challenges in interpreting and conveying their findings.

In systematic reviews, it is common to encounter publication bias. Due to the uncertain breadth and depth of our review's outcomes, we limited our search to peer-reviewed publications, amplifying this bias. It is noteworthy that other studies not subjected to this peer-review criterion might offer valuable insights. Therefore, we recommend that interested researchers and practitioners consult the tracks listed in Section 4.3 to obtain a broader perspective on the reviewed topic.

7 | CONCLUSION

In our systematic review, we analyzed 72 studies to explore the methods scholars have employed in multidimensional relevance estimation within the field of IR. The multidimensional nature of relevance is complex and diversely conceptualized across domains. This complexity, coupled with the variety of terminologies and methodologies, has presented challenges in standardizing definitions and operationalizations. To bring clarity, we proposed a structured formulation emphasizing clear definitions and transparent operational relationships between relevance factors. This approach promotes consistent future research. The recent advent of LLMs amplifies the timely significance of our review. With their advanced relevance labeling capabilities, LLMs offer potential solutions to challenges in creating benchmark collections for multidimensional relevance, a task traditionally reliant on human annotators. However, the success of LLMs relies on crafting precise prompts. Our review, especially the detailed definitions of relevance factors, can guide this prompt creation process. Moreover, LLMs′ may in future be able to estimate topical relevance scores. This development signifies a potential shift in IR, moving from models focused solely on topical relevance to those embracing multidimensional relevance. By considering user, task, and domain characteristics, such models mark a promising future direction, as they might offer a closer approximation to *user relevance*. In summary, our review sheds light on the complexities of multidimensional relevance, proposes a pathway for future research, and underscores the transformative potential of the domain due to the advancement of LLMs.

AUTHOR CONTRIBUTIONS

Georgios Peikos: Conceptualization (equal); data curation (equal); formal analysis (equal); methodology (equal); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Gabriella Pasi:** Conceptualization (supporting); methodology (supporting); supervision (lead); writing – review and editing (equal).

FUNDING INFORMATION

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860721.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

RELATED WIREs ARTICLES

A survey of Web crawlers for information retrieval

Text-based question answering from information retrieval and deep neural network perspectives: A survey

Soft clustering for information retrieval applications

ORCID

Georgios Peikos https://orcid.org/0000-0002-2862-8209

Gabriella Pasi https://orcid.org/0000-0002-6080-8170

ENDNOTES

[1] http://www.odp.org/homepage.php

[2] https://trec.nist.gov/, accessed on September 26, 2023.

[3] https://www.clef-initiative.eu/, accessed September 26, 2023.

[4] https://research.nii.ac.jp/ntcir/index-en.html, accessed September 26, 2023.

REFERENCES

Alsulmi, M., & Carterette, B. (2018). Improving medical search tasks using learning to rank. In *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1–8). IEEE.

Amancio, L., Dorneles, C. F., & Dalip, D. H. (2021). Recency and quality-based ranking question in cqas: A stack overflow case study. *Information Processing and Management*, *58*, 102552.

Arastoopoor, S. (2018). Domain-specific readability measures to improve information retrieval in the Persian language. *The Electronic Library*, *36*, 430–444.

Ashoori, E., & Lalmas, M. (2007). Using topic shifts for focussed access to xml repositories. In G. Amati, C. Carpineto, & G. Romano (Eds.), *Advances in Information Retrieval* (pp. 444–455). Springer.

Badache, I., & Boughanem, M. (2014). Social priors to estimate relevance of a resource. In *Proceedings of the 5th Information Interaction in Context Symposium, IIiX '14* (pp. 106–114). Association for Computing Machinery.

Balagopalan, A., Jacobs, A. Z., & Biega, A. J. (2023). The role of relevance in fair ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23* (pp. 2650–2660). Association for Computing Machinery.

Bambia, M., & Faiz, R. (2015). Frel: A freshness language model for optimizing real-time web search. In R. Silhavy, R. Senkerik, Z. K. Oplatkova, Z. Prokopova, & P. Silhavy (Eds.), *Intelligent systems in cybernetics and automation theory* (pp. 207–216). Springer International Publishing.

Banerjee, S., Upadhyay, R., Pasi, G., & Viviani, M. (2023). Summary in action: A trade-off between effectiveness and efficiency in multi-dimensional relevance estimation. In *IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT 2023, Venice, Italy, October 26–29, 2023 (pp. 119–126). IEEE.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, *34*(2–3), 219–236.

Bassani, E., & Pasi, G. (2021). A multi-representation re-ranking model for personalized product search. *Information Fusion*, *81*, 240–249.

Belkin, N. J. (2016). People, interacting with information1. In *ACM SIGIR Forum* (Vol. 49, pp. 13–27). ACM.

Bendersky, M., Croft, W. B., & Diao, Y. (2011). Quality-biased ranking of web documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11* (pp. 95–104). Association for Computing Machinery.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3), 151.

Bouidghaghen, O., Tamine-Lechani, L., Pasi, G., Cabanac, G., Boughanem, M., & da Costa Pereira, C. (2011). Prioritized aggregation of multiple context dimensions in mobile IR. In M. V. M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, & H. Khelalfa (Eds.), *Information retrieval technology* (pp. 169–180). Springer Berlin Heidelberg.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Burges, C. J. (2010). From RankNet to LambdaRank to LambdaMart: An overview. *Learning*, *11*(23–581), 81.

Carmel, D., Haramaty, E., Lazerson, A., & Lewin-Eytan, L. (2020). Multi-objective ranking optimization for product search using stochastic label aggregation. In *Proceedings of the Web Conference 2020, WWW '20* (pp. 373–383). Association for Computing Machinery.

Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Brookline Books.

Chapelle, O., & Chang, Y. (2011). Yahoo! Learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge* (pp. 1–24). PMLR.

Chenlo, J. M., Parapar, J., Losada, D. E., & Santos, J. (2015). Finding a needle in the blogosphere: An information fusion approach for blog distillation search. *Information Fusion*, *23*, 58–68.

Chu, H. (2011). Factors affecting relevance judgment: A report from TREC legal track. *Journal of Documentation*, *67*(2), 264–278.

Clarke, C. L. A., Rizvi, S., Smucker, M. D., Maistro, M., & Zuccon, G. (2020). Overview of the TREC 2020 health misinformation track. In E. M. Voorhees & A. Ellis (Eds.), *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event*, Gaithersburg, Maryland, USA, November 16–20, 2020, Volume 1266 of NIST Special Publication. National Institute of Standards and Technology (NIST).

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 403–412). USA. Association for Computing Machinery.

Cool, C., Belkin, N., Frieder, O., & Kantor, P. (1993). Characteristics of text affecting relevance judgments. In *National Online Meeting* (Vol. 14, p. 77). Learned Information (Europe) Ltd.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Cooper, W. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), 19–37.

Cosijn, E. (2009). Relevance judgments and measurements. In *Encyclopedia of library and information sciences* (pp. 4512–4519). CRC Press.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550.

Craswell, N., Robertson, S., Zaragoza, H., & Taylor, M. (2005). Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05* (pp. 416–423). Association for Computing Machinery.

Crestani, F., Mizzaro, S., & Scagnetto, I. (2017). *Mobile information retrieval*. Springer.

da Costa Pereira, C., Dragoni, M., & Pasi, G. (2009). Multidimensional relevance: A new aggregation criterion. In M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.), *Advances in information retrieval* (pp. 264–275). Springer.

da Costa Pereira, C., Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing and Management*, *48*(2), 340–357.

Dai, N., Shokouhi, M., & Davison, B. D. (2011). Learning to rank for freshness and relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11* (pp. 95–104). Association for Computing Machinery.

Daoud, M., Daoud, M., & Huang, J. X. (2013). Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the Association for Information Science and Technology*, *64*, 190–212.

Dumitrescu, A., & Santini, S. (2021). Full coverage of a reader's interests in context-based information filtering. *Journal of the Association for Information Science and Technology*, *72*(8), 1011–1027.

Eickhoff, C., & de Vries, A. P. (2014). Modelling complex relevance spaces with copulas. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14* (pp. 1831–1834). Association for Computing Machinery.

Eickhoff, C., de Vries, A. P., & Collins-Thompson, K. (2013). Copulas for information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13* (pp. 663–672). Association for Computing Machinery.

Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J., & SanJuan, E. (2017). CLEF 2017 microblog cultural contextualization lab overview. In G. J. F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction—Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017*, Dublin, Ireland, September 11–14, 2017, Volume 10456 of Lecture Notes in Computer Science (pp. 304–314). Springer.

Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., & Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '23* (pp. 39–50). Association for Computing Machinery.

Farah, M., & Vanderpooten, D. (2008). An outranking approach for information retrieval. *Information Retrieval*, *11*, 315–334.

Feng, J., Li, H., Huang, M., Liu, S., Ou, W., Wang, Z., & Zhu, X. (2018). Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning. In *Proceedings of the 2018 World Wide Web Conference, WWW '18*, Republic and Canton of Geneva, CHE (pp. 1939–1948). International World Wide Web Conferences Steering Committee.

Fernández-Pichel, M., Losada, D. E., & Pichel, J. C. (2022). A multistage retrieval system for health-related misinformation detection. *Engineering Applications of Artificial Intelligence*, *115*, 105211.

Frei, J., & Kramer, F. (2023). Annotated dataset creation through large language models for non-English medical NLP. *Journal of Biomedical Informatics*, *145*, 104478.

Gerani, S., Carman, M. J., & Crestani, F. (2010). Proximity-based opinion retrieval. In F. Crestani, S. Marchand-Maillet, H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010*, Geneva, Switzerland, July 19–23, 2010 (pp. 403–410). ACM.

Gerani, S., Zhai, C., & Crestani, F. (2012). Score transformation in linear combination for multi-criteria relevance ranking. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, & F. Silvestri (Eds.), *Advances in information retrieval*, Berlin, Heidelberg (pp. 256–267). Springer.

Giachanou, A., Harvey, M., & Crestani, F. (2016). Topic-specific stylistic variations for opinion retrieval on Twitter. In N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, & G. Silvello (Eds.), *Advances in information retrieval—Proceedings of the 38th European Conference on IR Research, ECIR 2016*, Padua, Italy, March 20–23, 2016. Volume 9626 of Lecture Notes in Computer Science (pp. 466–478). Springer.

Goffman, W., & Newill, V. A. (1966). A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval*, *3*(1), 19–25.

Huang, J. X., He, B., & Zhao, J. (2018). Mining authoritative and topical evidence from the blogosphere for improving opinion retrieval. *Information Systems*, *78*, 199–213.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context* (Vol. 18). Springer Science & Business Media.

Jabeur, L. B., Tamine, L., & Boughanem, M. (2012). Featured tweet search: Modeling time and social influence for microblog retrieval. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 166–173). IEEE Computer Society.

Jomsri, P., & Prangchumpol, D. (2015). A hybrid model ranking search result for research paper searching on social bookmarking. In *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)* (pp. 38–43). IEEE.

Kang, C., Wang, X., Chang, Y., & Tseng, B. (2012). Learning to rank with multi-aspect relevance for vertical search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12* (pp. 453–462). Association for Computing Machinery.

Karmaker Santu, S. K., Sondhi, P., & Zhai, C. (2017). On application of learning to rank for e-commerce search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17* (pp. 475–484). Association for Computing Machinery.

Klare, G. R. (2000). The measurement of readability: Useful information for communicators. *ACM Journal of Computer Documentation*, *24*(3), 107–121.

Komatsuda, T., Keyaki, A., & Miyazaki, J. (2016). A score fusion method using a mixture copula. In S. Hartmann & H. Ma (Eds.), *Database and expert systems applications* (pp. 216–232). Springer International Publishing.

Li, J., Zhang, P., Song, D., & Wu, Y. (2017). Understanding an enriched multidimensional user relevance model by analyzing query logs. *Journal of the Association for Information Science and Technology*, *68*, 2743–2754.

Li, X., Mao, J., Ma, W., Liu, Y., Zhang, M., Ma, S., Wang, Z., & He, X. (2021). Topic-enhanced knowledge-aware retrieval model for diverse relevance estimation. In *Proceedings of the Web Conference 2021, WWW '21* (pp. 756–767). Association for Computing Machinery.

Lioma, C., Larsen, B., Lu, W., & Huang, Y. (2016). A study of factuality, objectivity and relevance: Three desiderata in large-scale information retrieval? In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT '16* (pp. 107–117). Association for Computing Machinery.

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, *58*(3), 102522.

Ma, Y., Wu, Y., Ai, Q., Liu, Y., Shao, Y., Zhang, M., & Ma, S. (2023). Incorporating structural information into legal case retrieval. *ACM Transactions on Information Systems*, *42*, 1–28.

Madisetty, S., & Desarkar, M. S. (2022). A reranking-based tweet retrieval approach for planned events. *World Wide Web*, *25*(1), 23–47.

Mandayam Comar, P., & Sengamedu, S. H. (2017). Intent based relevance estimation from click logs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17* (pp. 59–66). Association for Computing Machinery.

McGregor, M., Azzopardi, L., & Halvey, M. (2023). A systematic review of cost, effort, and load research in information search and retrieval, 1972–2020. *ACM Transactions on Information Systems*, *42*(1), 1–29.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, *10*(3), 303–320.

Moulahi, B., Moulahi, B., Moulahi, B., Tamine, L., & Yahia, S. B. (2014). Iaggregator: Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, *65*, 2062–2083.

Moulahi, B., Tamine, L., & Yahia, S. B. (2014). Toward a personalized approach for combining document relevance estimates. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User modeling, adaptation, and personalization* (pp. 158–170). Springer International Publishing.

Omidvar-Tehrani, B., Personnaz, A., & Amer-Yahia, S. (2022). Guided text-based item exploration. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22* (pp. 3410–3420). Association for Computing Machinery.

Palacio, D., Cabanac, G., Sallaberry, C., & Hubert, G. (2010). On the evaluation of geographic information retrieval systems: Evaluation framework and case study. *International Journal on Digital Libraries*, *11*, 91–109.

Palotti, J., Zuccon, G., & Hanbury, A. (2019). Consumer health search on the web: Study of web page understandability and its integration in ranking algorithms. *Journal of Medical Internet Research*, *21*, e10986.

Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science*, *45*, 135–141.

Pasi, G., & Viviani, M. (2018). Application of aggregation operators to assess the credibility of user-generated content in social media. In *Information processing and management of uncertainty in knowledge-based systems. Theory and foundations: Proceedings of the 17th International Conference, IPMU 2018*, Cádiz, Spain, June 11–15, 2018, Part I 17 (pp. 342–353). Springer.

Putri, D. G. P., Viviani, M., & Pasi, G. (2020). Social search and task-related relevance dimensions in microblogging sites. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social informatics* (pp. 297–311). Springer International Publishing.

Putri, D. G. P., Viviani, M., & Pasi, G. (2021). A multi-task learning model for multidimensional relevance assessment. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction* (pp. 103–115). Springer International Publishing.

Qu, J., Arguello, J., & Wang, Y. (2020). Towards explainable retrieval models for precision medicine literature search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20* (pp. 1593–1596). Association for Computing Machinery.

Qu, J., Arguello, J., & Wang, Y. (2021). A deep analysis of an explainable retrieval model for precision medicine literature search. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in information retrieval* (pp. 544–557). Springer International Publishing.

Ravikumar, S., Talamadupula, K., Balakrishnan, R., & Kambhampati, S. (2013). Raprop: Ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13* (pp. 2345–2350). Association for Computing Machinery.

Rekabsaz, N., & Lupu, M. (2014). A real-world framework for translator as expert retrieval. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, & E. Toms (Eds.), *Information access evaluation. Multilinguality, multimodality, and interaction* (pp. 141–152). Springer International Publishing.

Rinaldi, A. M. (2009). An ontology-driven approach for semantic information retrieval on the web. *ACM Transactions on Internet Technology*, *9*(3), 1–24.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In D. K. Harman (Ed.), *Proceedings of the Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2–4, 1994, volume 500–225 of NIST Special Publication* (pp. 109–126). National Institute of Standards and Technology (NIST).

Sahraoui, A. K., & Faiz, R. (2017). Time sensitivity for personalized search. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 585–592). IEEE.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the annual meeting-american society for information science* (Vol. 34, pp. 313–327). Learned Information (Europe) Ltd.

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.

Saracevic, T. (2016). *The notion of relevance in information science: Everybody knows what relevance is. But, what is it really*? Morgan & Claypool Publishers.

Sasaki, Y., Komatsuda, T., Keyaki, A., & Miyazaki, J. (2016). A new readability measure for web documents and its evaluation on an effective web search engine. In *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services, iiWAS '16* (pp. 355–362). Association for Computing Machinery.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, *26*(6), 755–776.

Shajalal, M., & Aono, M. (2020). Coverage-based query subtopic diversification leveraging semantic relevance. *Knowledge and Information Systems*, *62*, 2873–2891.

Shraga, R., Roitman, H., Feigenblat, G., & Cannim, M. (2020). Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20* (pp. 1399–1408). Association for Computing Machinery.

Sieg, A., Mobasher, B., & Burke, R. (2007). Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07* (pp. 525–534). Association for Computing Machinery.

Silva, R. F., Roy, C. K., Rahman, M. M., Schneider, K. A., Paixao, K., & de Almeida Maia, M. (2019). Recommending comprehensive solutions for programming tasks by mining crowd knowledge. In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)* (pp. 358–368). IEEE/ACM.

Singh, V., & Dave, M. (2019). Improving result diversity using query term proximity in exploratory search. In S. Madria, P. Fournier-Viger, S. Chaudhary, & P. K. Reddy (Eds.), *Big data analytics* (pp. 67–87). Springer International Publishing.

Sun, Y., Zhang, Y., Gwizdka, J., & Trace, C. B. (2019). Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *Journal of Medical Internet Research*, *21*(5), e12522.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, *56*(4), 389–398.

Tamine, L., & Chouquet, C. (2017). On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing and Management*, *53*(2), 332–350.

Tamine, L., Jabeur, L. B., & Bahsoun, W. (2011). *On using social context to model information retrieval and collaboration in scientific research community* (pp. 133–155). Springer Berlin Heidelberg.

Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V., & de Rijke, M. (2021). Challenges and research opportunities in ecommerce search and recommendations. In *ACM SIGIR Forum* (Vol. 54, pp. 1–23). ACM.

Upadhyay, R., Pasi, G., & Viviani, M. (2022). An unsupervised approach to genuine health information retrieval based on scientific evidence. In R. Chbeir, H. Huang, F. Silvestri, Y. Manolopoulos, & Y. Zhang (Eds.), *Web information systems engineering—WISE 2022* (pp. 119–135). Springer International Publishing.

Upadhyay, R., Pasi, G., & Viviani, M. (2023). A passage retrieval transformer-based re-ranking model for truthful consumer health search. In D. Koutra, C. Plant, M. Gomez-Rodriguez, E. Baralis, & F. Bonchi (Eds.), *Machine learning and knowledge discovery in databases: Research track—European Conference, ECML PKDD 2023*, Turin, Italy, September 18–22, 2023, Proceedings, Part I, volume 14169 of Lecture Notes in Computer Science (pp. 355–371). Springer.

Uprety, S., Su, Y., Song, D., & Li, J. (2018). Modeling multidimensional user relevance in ir using vector spaces. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18* (pp. 993–996). Association for Computing Machinery.

Usta, A., Altingovde, I. S., Ozcan, R., & Ulusoy, Ö. (2021). *Learning to rank for educational search engines*. IEEE Transactions on Learning Technologies.

Vakkari, P. (2020). The usefulness of search results: A systematization of types and predictors. In *Proceedings of the 2020 conference on human information interaction and retrieval, CHIIR '20* (pp. 243–252). Association for Computing Machinery.

van Doorn, J., Odijk, D., Roijers, D. M., & de Rijke, M. (2016). Balancing relevance criteria through multi-objective optimization. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16* (pp. 769–772). Association for Computing Machinery.

van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artifical Intelligence and Law*, *25*(1), 65–87.

Vargas, S., Castells, P., & Vallet, D. (2012). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12* (pp. 75–84). USA. Association for Computing Machinery.

Vickery, B. C. (1959a). The structure of information retrieval systems. *Proceedings of the International Conference on Scientific Information*, *2*, 1275–1290.

Vickery, B. C. (1959b). Subject analysis for information retrieval. *Proceedings of the International Conference on Scientific Information*, *2*, 855–865.

Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W., & Li, Y. (2005). Detecting dominant locations from search queries. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, August 15–19, 2005 (pp. 424–431). ACM.

Wiebe, J., & Riloff, E. (2011). Finding mutual benefit between subjectivity analysis and information extraction. *IEEE Transactions on Affective Computing*, *2*(4), 175–191.

Wiggers, G., Verberne, S., van Loon, W., & Zwenne, G.-J. (2023). Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance. *Journal of the Association for Information Science and Technology*, *74*, 1010–1025.

Wiggers, G., Verberne, S., & Zwenne, G. (2018). Exploration of intrinsic relevance judgments by legal professionals in information retrieval systems. In *Proceedings of the 17th Dutch-Belgian Information Retrieval Workshop* (pp. 5–8). Leiden University, Scholarly Publications.

Xu, B., Lin, H., Lin, Y., Ma, Y., Yang, L., Wang, J., & Yang, Z. (2016). Improve biomedical information retrieval using modified learning to rank methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(6), 1797–1809.

Xu, Y., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973.

Yan, M., Wen, Y., Shi, Q., Tian, X., & Zhao, L. (2022). A multimodal retrieval and ranking method for scientific documents based on HFS and XLNet. *Scientific Programming*, *2022*, 1–11.

Yang, H., Gupta, P., Fernández Galán, R., Bu, D., & Jia, D. (2021). Seasonal relevance in e-commerce search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21* (pp. 4293–4301). Association for Computing Machinery.

Yulianti, E., Chen, R.-C., Scholer, F., Croft, W. B., & Sanderson, M. (2018). Ranking documents by answer-passage quality. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18* (pp. 335–344). Association for Computing Machinery.

Zhang, Q., & Youssef, A. (2014). An approach to math-similarity search. In S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, & J. Urban (Eds.), *Intelligent computer mathematics* (pp. 404–418). Springer International Publishing.

Zhang, W., Song, D., Zhang, P., Zhao, X., & Hou, Y. (2015). A sequential latent topic-based readability model for domain-specific information retrieval. In G. Zuccon, S. Geva, H. Joho, F. Scholer, A. Sun, & P. Zhang (Eds.), *Information retrieval technology* (pp. 241–252). Springer International Publishing.

Zhuang, H., Wang, X., Bendersky, M., Grushetsky, A., Wu, Y., Mitrichev, P., Sterling, E., Bell, N., Ravina, W., & Qian, H. (2021). Interpretable ranking with generalized additive models. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21* (pp. 499–507). Association for Computing Machinery.

Znaidi, E., Tamine, L., & Latiri, C. (2016). Aggregating semantic information nuggets for answering clinical queries. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, SAC '16* (pp. 1041–1047). Association for Computing Machinery.

How to cite this article: Peikos, G., & Pasi, G. (2024). A systematic review of multidimensional relevance estimation in information retrieval. *WIREs Data Mining and Knowledge Discovery*, e1541. https://doi.org/10.1002/widm.1541

APPENDIX A: SYSTEMATIC REVIEW METHOD

This first part of the appendix provides all the details related to the collection, selection, and coding of the acquired publications. The details presented here clarify the gathering process and aid in understanding the scope and breadth of the literature reviewed, ensuring the transparency of the employed research methodology. As the first step (i.e., the research questions) of our systematic literature review has been presented, we proceed with the four remaining steps.

Step 2: Inclusion and exclusion criteria

The aim of this step was to establish and evaluate the inclusion and exclusion criteria that were utilized to systematically select and reject articles for review. The development of the inclusion and exclusion criteria commenced by

TABLE A1 List of selection criteria.

Inclusion/exclusion criteria
Including studies focused on text retrieval
Including empirical studies that utilize a minimum of two relevance factors
Including scholarly publications subject to peer review
Including both full-length research articles and short papers
Excluding studies solely focused on operationalizing a relevance factor
Sources are confined to journals, conference proceedings, and workshops
No specific time frame
Studies must be written in English

compiling criteria that align with the target study type: multidimensional relevance estimation in IR. Although the initial list of criteria was seen as provisional and subject to refinement throughout the review process (i.e., after processing 10% of total included articles), no further adaptations to the criteria were implemented. Table A1 presents a comprehensive list of the final criteria.

This review exclusively included studies focusing on text retrieval systems (i.e., document retrieval), as studies involving other types of information objects (e.g., audio, video) would significantly expand the scope of the study. This review encompassed empirical studies (i.e., use experimental methods) that utilized a minimum of two relevance factors for document ranking, with topical relevance being one of those factors. Consequently, we omitted studies that employed neural models for document re-ranking, as these studies rely solely on topical relevance signals. In this review, we excluded studies in which researchers solely operationalized a relevance factor, without utilizing it to estimate multidimensional relevance and perform document ranking. We applied this exclusion criterion since the studies primarily aimed to predict a single score for a relevance factor rather than estimate multidimensional relevance. Our review specifically investigated how relevance factors have been operationalized only when they were utilized for retrieval.

To ensure the selection of higher quality articles, the inclusion criteria were restricted to scholarly publications that had undergone peer review. Consequently, sources were confined to journals, conference proceedings, and workshops, encompassing both full-length research articles and short papers. This criterion lead to the potential exclusion of essential initiatives' proceedings (see Table A2). Nonetheless, several of these papers were still included as they were later published in peer-reviewed journal or conferences. Moreover, our systematic review reports on benchmark collections that are often associated with the aforementioned initiatives, providing a reference point for interested readers. Ultimately, to capture the complete scope of relevant articles, a specific time frame was not imposed, and all studies included in the review were required to be written in English.

Step 3: Search strategy and paper selection

We used the inclusion and exclusion criteria outlined previously to acquire publications on multidimensional relevance estimation in IR. These were obtained through searches across multiple research publication search engines and databases. The process of searching for potentially relevant articles for this review consists of the following steps, as shown in Figure A1.

Similarly to the systematic review conducted by McGregor et al. (2023), we initiated the search process by searching within the selected literature databases (journals and conferences/workshops) shown in Table A3. To facilitate the database search, we created the following query: (*multidimensional relevance OR relevance factors OR relevance dimensions OR relevance aspects OR multi aspect relevance*) *AND* (*information retrieval*). For the majority of the resources, the search was refined to "title" and "abstract" search. However, in cases that this was not feasible, we conducted the search using the "full-text" option. To avoid missing relevant articles, we additionally conducted searches in Google Scholar, Springer Link, ACM Digital Library, IEEE Xplore, and Science Direct, similarly to previous studies (Liu, 2021; Vakkari, 2020). These searches also utilize the same query, with slight modifications tailored to their specific requirements. We tried different combinations of the aforementioned keywords, aiming to cover most, if not all, of the relevant research for further analysis. Following the aforementioned search process, a total of 1387 studies have been identified. Those articles have been manually screened by reviewing their title and abstracts to determine their relevance to this study. At this point, we were interested in reducing the initial document pool to include those focused on document

TABLE A2 List of initiatives.

Initiatives	
CLEF	Conference and Labs of the Evaluation Forum
TREC	Text Retrieval Conference
FIRE	Forum for Information Retrieval Evaluation
INEX	Initiative for the Evaluation of XML Retrieval
NTCIR	NII Testbeds and Community for Information Access Research

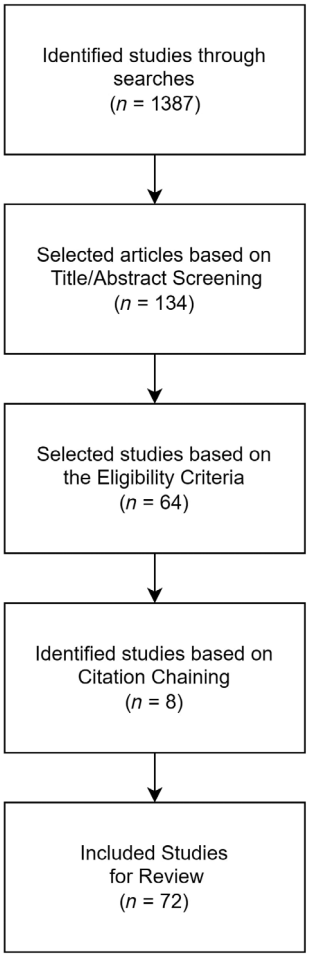

FIGURE A1 Overview of the followed search process.

retrieval and excluding those studies that solely estimate a score to be associated with a relevance factor without using it for ranking. As a result, a total of 134 studies have been selected for further examination. These studies have been evaluated based on the whole set of inclusion/exclusion criteria listed in Table A1, and a total of 64 studies have been identified as eligible for this study. The majority of the papers have been excluded because they were not focused on text retrieval. Finally, similarly to Liu (2021) and McGregor et al. (2023), we performed a forward and backward citation chaining on the final pool of the 64 eligible studies and, 8 additional studies were included for review.

Steps 4 and 5: Coding scheme and paper synthesis

The categories for coding and analysis were designed in accordance with the research questions (RQs) we aimed to address. The employed coding scheme consists of general information related to publication characteristics such as authors' affiliations, publication venues and year. The purpose was to provide insights into the distribution of research across different areas and over time. Aiming to address RQ1, we coded studies based on the knowledge domain exploited in their experimental evaluations, the employed relevance factors, and the exploited approach to aggregate the relevance scores and rank the documents. The identified aggregation approaches were categorized to data-driven, model-driven, or other (for those that did not clearly fit in the other categories). The second research question aimed at providing insights related to the investigated relevance factors. To that aim, we highlighted similarities and differences between the definitions and operationalizations of the identified relevance factors across studies and knowledge domains. By comparing and contrasting the identified studies based on how they exploit the associated relevance factors, we obtained a clearer understanding regarding conceptual and experimental differences. Finally, regarding the third research question, the included studies have been coded based on their employed benchmark collection, which

TABLE A3 Sources examined in database and other searches.

Journals
Information Processing and Management (IP&M)
Journal of the Association for Information Science and Technology (JASIS&T)
International Journal on Digital Libraries
Information Retrieval Journal (IRJ)
Journal of Information Science
Journal of Documentation
ACM Transactions on Information Systems (TOIS)
Conferences/workshops
ACM/IEEE Joint Conference on Digital Libraries (JCDL)
European Conference on Digital Libraries (ECDL)
European Conference on Information Retrieval (ECIR)
ACM International Conference on Information and Knowledge Management (CIKM)
Proceedings of the Association of Information Science and Technology (ASIS&T)
ACM Special Interest Group on Information Retrieval Conference (SIGIR)
ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)
Information Interaction in Context Conference (IIiX)
ACM International Conference on Web Search and Data Mining (WSDM)
International Conference on the Theory of Information Retrieval (ICTIR)
ACM Conference on Recommender Systems Conference (RecSys)
Other sources
Google Scholar
Springer Link
ACM Digital Library
IEEE Xplore
Science Direct

have been further analyzed regarding their characteristics. As a result, we obtained insights regarding the available benchmark collections that can be used to investigate multidimensional relevance models.

Following the coding schema as described above, we were able to identify commonalities and differences regarding multidimensional relevance estimation, across knowledge domains and search tasks. Through this systematic review, we obtained a better understanding of the limitations and potentials of exploiting relevance as multidimensional concept in IR. Our analysis allows to compare studies in terms of the aggregation methods, the application domains, and the relevance factors (definition, operationalization). Synthesizing them based on the application domain, we draw insights regarding the definition and operationalization of the employed relevance factors. Synthesizing them based on the relevance factors, we investigate how these factors are exploited across domains. Finally, by analyzing their datasets, we draw insights regarding their similarities and differences and we highlight future necessities.

APPENDIX B: OVERALL PUBLICATION CHARACTERISTICS

This section presents key characteristics of the publications under study. We explore a multifaceted view of the research landscape in multidimensional relevance estimation, by examining the publications based on: their geographical distribution; the collaborative efforts between industry and academia highlighting synergies; the diversity in types of publication venues; and the temporal distribution that offers insights into the evolution of research in this domain.

Central to our review, we identified 200 researchers who have significantly contributed to the literature on this subject. These researchers represent a wide spectrum of expertise, originating from varied academic and professional backgrounds. Our review reveals a diverse geographical distribution of research on multidimensional relevance estimation. A detailed representation of the number of papers per country, based on authors' affiliations, is provided in Figure B1. As illustrated in the figure, the United States leads in contributions with 16 studies, closely followed by China, France, and Italy. Similarly, several European countries have shown significant contributions, with Italy, United Kingdom, France, Spain, and the Netherlands collectively accounting for 35 studies. Notably, Tunisia stands out in the North African region with five contributions, while Asia's presence is also marked by contributions from countries such as China, Japan, India, and South Korea. The global map illustrating this geographic distribution provides a comprehensive snapshot of the worldwide research landscape in the examined area, highlighting the strong collaboration among researchers.

A noteworthy observation from our review is the synergy between academia and industry, as shown in Figure B2.

We quantified the collaborations and found 12 of the included publications showcasing a partnership between academia and industry. A total of 7 publications is being authored by researchers working in industry. Several studies were conducted by major corporations in the field such as Microsoft (Collins-Thompson et al., 2011; Craswell et al., 2005), Google (Zhuang et al., 2021), Yahoo (Kang et al., 2012), and Amazon (Carmel et al., 2020; Mandayam Comar & Sengamedu, 2017; Yang et al., 2021), as well as other companies collaborating with universities to address information retrieval tasks in domain-specific search (Sasaki et al., 2016; Wiggers et al., 2023). Such collaborations are indicative of the practical applications and real-world significance of estimating relevance by considering several factors that affect it under specific contextual situations.

Regarding the distribution of publication venues over time, this is illustrated in Figure B3. As we previously discussed, the idea of considering relevance as a multidimensional concept is rooted in the origins of information search systems (Saracevic, 2007). Contributions by researchers such as Goffman and Newill (1966), Cooper (1971), and Mizzaro (1998), among many others, lead to a shift toward recognizing its dynamic and multidimensional nature. Following this recognition, several researchers conducted user studies to identify contributing relevance factors, with key

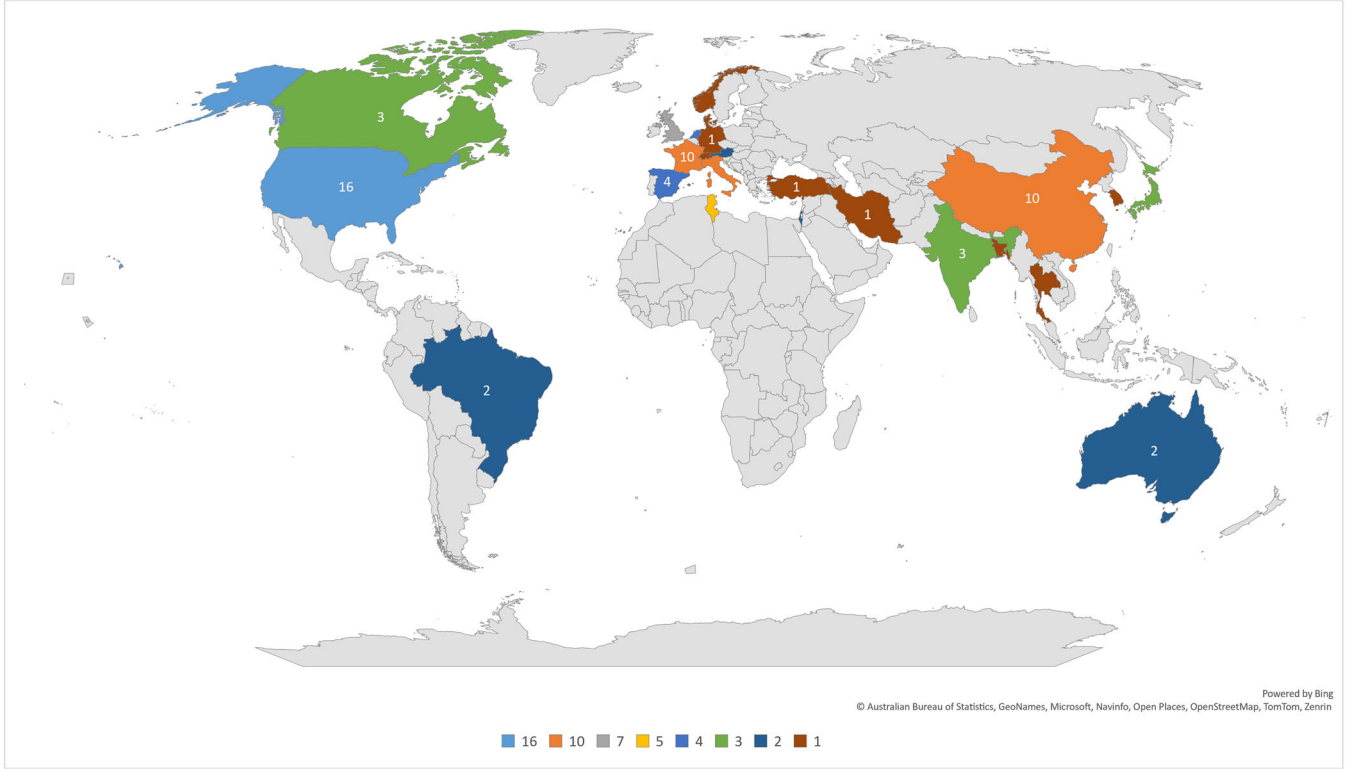

FIGURE B1 Map illustrating the number of publications on multidimensional relevance estimation by country. The geographic location is determined by the authors' affiliations and not their nationalities.

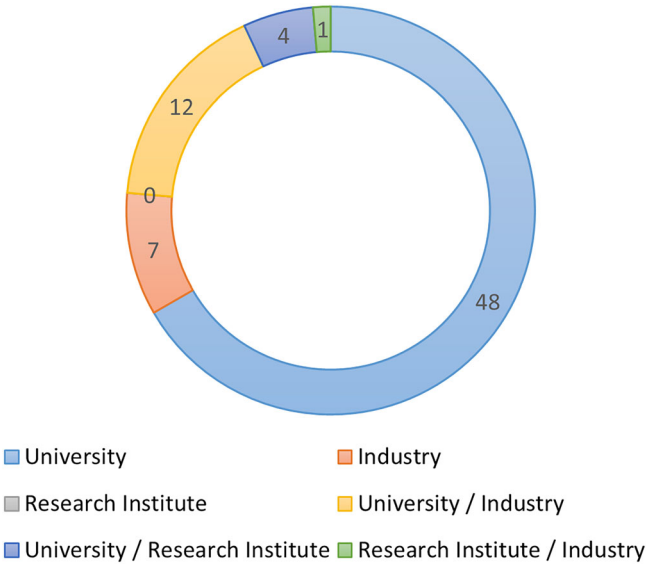

FIGURE B2 A representation of synergies between universities, research institutions, and industry in the studied literature.

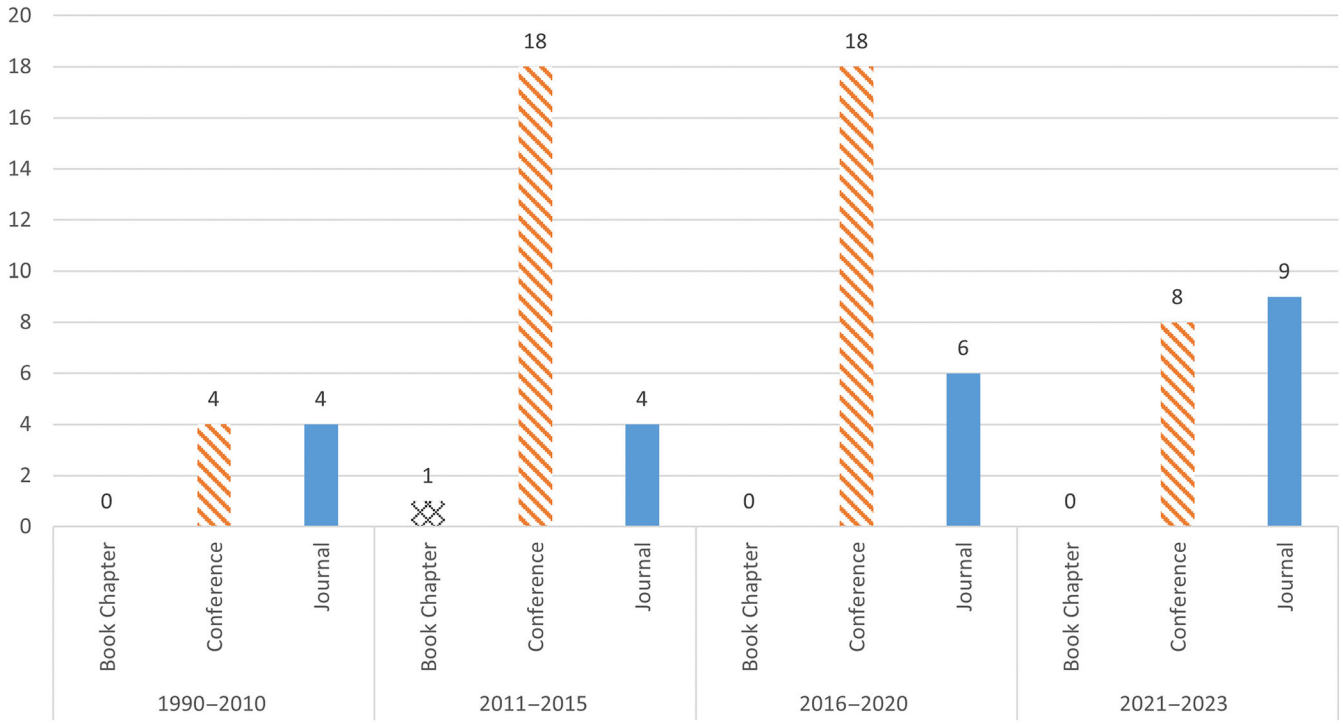

FIGURE B3 Distribution of papers across venue types over time intervals.

studies being from Barry and Schamber (1998), Cool et al. (1993), Xu and Chen (2006), among others. Subsequently, experimental evaluations were pursued by multiple scholars (Ashoori & Lalmas, 2007; Brin & Page, 1998; Craswell et al., 2005; Farah & Vanderpooten, 2008; Sieg et al., 2007), with the most notable contribution that utilized multiple relevance factors for ranking is the integration of the PageRank algorithm in commercial web search (Brin & Page, 1998). In the following years (2011–2020), one can observe a consistent trend in publications, with both the periods 2011–2015 and 2016–2020 showing nearly identical numbers of conference and journal publications. This suggests a stable and sustained research interest in the topic throughout the decade. From 2021 to 2023, there one can observe an upward trend in journal publications. However, this observation might not provide a full comparison with

the previous years for two reasons: (1) the time span under consideration is shorter, and (2) several of the identified publications in 2023 have not been peer reviewed and have been excluded from our review.

Among the 28 identified conferences, the ACM SIGIR Conference on Research and Development in Information Retrieval stands out as a primary venue, hosting 10 out of the 72 surveyed papers, followed by the Conference on Information and Knowledge Management (CIKM) with 6 publications. There are 18 distinct Journals, from which the Journal of the Association for Information Science and Technology emerges as a leading venue with 3 publications, followed by journals such as Information Fusion and the Information Processing and Management Journal that have 2 publications.