AND DATA SCIENCES
BULLETIN OF COMPUTER

RC-Rank: A User-Oriented Multi-Criteria Ranking Framework for Scientific Data Search

Mobeen Akhter and Nazish Noreen

INTI International University, Malaysia

Abstract

As scientific practice becomes increasingly data-intensive, researchers depend on repositories and portals to locate datasets that are not only topically related to their tasks but also trustworthy, accessible, and reusable. Prior work on scientific data relevance has shown that users employ multiple criteria—including topicality, authority, quality, accessibility, and utility—in structured sequences when deciding whether a dataset is relevant. However, current dataset search engines largely rely on keyword-based ranking over metadata and rarely operationalize such user-oriented models in their ranking algorithms. This paper proposes RC-Rank, a user-oriented multi-criteria ranking framework for scientific data search that explicitly encodes relevance criteria into the ranking function. RC-Rank groups features into criterion-specific channels, defines a principled aggregation scheme over these channels, and learns criterion weights and feature parameters from user interaction data. We outline a practical feature design for common scientific data portals, present a criterion-aware scoring formulation, and illustrate the behavior of RC-Rank on a synthetic case study. We further describe an evaluation protocol combining offline learning-to-rank experiments with user studies to assess both retrieval effectiveness and perceived alignment with human relevance judgments. The framework bridges conceptual models of user relevance with deployable ranking algorithms, and provides a basis for building cognition-friendly dataset search systems.

Keywords: dataset search, scientific data, ranking algorithms, relevance criteria, multi-criteria decision making

1. Introduction

The exponential growth of scientific data has led to the proliferation of repositories, portals, and registries in virtually every domain, from biomedicine and environmental science to social science and astronomy. Large disciplinary infrastructures such as GenBank, ESA's Climate Data Store, and ICPSR, as well as national and institutional open data portals, now expose thousands or even millions of datasets for public use. At the same time, cross-repository services such as Google Dataset Search and domain-specific aggregators integrate metadata from heterogeneous sources and promise a single entry point to this increasingly complex ecosystem [1]. As a result, researchers, policy makers, and practitioners increasingly rely on these infrastructures to discover datasets for reuse, integration, and secondary analysis, rather than collecting all data themselves.

Corresponding author: Mobeen Akhter (mobeen2398@gmail.com)

BCDS 2025, Vol. 6(1), 61-80. DOI: 10.71448/bcds2561-4

This growth in supply has transformed dataset discovery into a central bottleneck for data-intensive science. Decades of information retrieval research have shown that relevance is multi-dimensional and situational [2], and recent work on research data confirms that similar complexity exists for data: searching for a dataset is not merely a matter of matching keywords in a title, but a process of interpreting metadata, documentation, provenance, and community practices [3, 4]. In contrast to traditional document search, the decision to reuse a dataset depends on whether it can be meaningfully integrated into an analysis pipeline, whether its methods are credible, whether licensing permits the intended use, and whether its temporal, spatial, or population coverage matches the task at hand.

Empirical studies of scientific data users have begun to unpack how these decisions are made in practice. Liu et al. show that relevance judgments for scientific data are multi-faceted and process-oriented: users typically begin by assessing topicality, then examine indicators of reliability such as perceived data quality and authority, and finally evaluate whether a candidate dataset is useful for their specific analytical task [5]. Similar patterns emerge in data sensemaking and data search diaries, where users report moving iteratively between judging what a dataset is "about", whether they can trust it, and whether it fits their intended use [4, 6]. These findings imply that ranking algorithms based purely on lexical similarity or simple metadata filters may diverge substantially from user judgments, particularly in specialized or high-stakes settings where choosing an inappropriate dataset carries a real cost.

In parallel, the dataset search literature has emphasized that current production systems—including institutional repositories, open data portals, and vertical search engines—continue to rely predominantly on keyword search over metadata with basic faceted filters [1]. This architecture scales well and is easy to deploy, but it often ignores richer signals about dataset trustworthiness, reuse history, and FAIR characteristics that are known to matter greatly to data users [8]. Analyses of query logs and user behavior on open data portals show that users struggle with vague or under-specified metadata, iterate through trial-and-error queries, and frequently resort to indirect strategies such as searching the literature for references to datasets or contacting colleagues for recommendations [3, 9]. Despite this evidence, the ranking functions that order search results in many portals still treat topical matching as the primary signal, with only modest adjustments for timeliness or popularity.

This paper argues that the gap between empirically grounded models of data relevance and operational ranking algorithms is now a central bottleneck for effective scientific data search. On the one hand, we have increasingly nuanced accounts of how data users think about relevance, including the roles of topicality, authority, quality, accessibility, and task-specific utility [5, 7]. On the other hand, the ranking functions deployed in real systems are rarely informed by these models and, in many cases, remain opaque even to repository operators. Bridging this gap requires a ranking formulation that (i) explicitly represents the criteria that users employ, (ii) makes the contribution of each criterion auditable and tunable, and (iii) can still be optimized using scalable learning-to-rank machinery where interaction data are available [10].

Our goal is therefore not to introduce a new cognitive theory of relevance, but to operationalize existing empirical insights in a form that can be embedded into contemporary dataset search architectures. We propose RC-Rank, a criterion-aware ranking framework that decomposes the ranking function into criterion-specific scoring components and an interpretable aggregation layer. The framework makes explicit design assumptions (what metadata and derived statistics are available, what is meant by a criterion score, and how scores are normalized) and provides a concrete feature blueprint

for typical portals. We also include an illustrative case study that demonstrates how different criterion weight configurations induce different, interpretable ranking trade-offs, and we specify an evaluation protocol that enables rigorous offline and user-centered validation. Importantly, the claims we make are bounded by the evidence presented: the contribution of this paper is the framework, its formalization, and a fully specified methodology for evaluating it, rather than an empirical claim of superiority on a particular deployed portal.

1.1. Problem statement

We consider the standard setting of a scientific data portal that indexes a collection of datasets $\mathcal{D}$, each described by a metadata record M_d and possibly additional derived statistics S_d that capture aspects such as usage, basic data profiling results, or provenance. Given a user query q expressed as a set of keywords, a short natural-language description, or a task template (for example, "daily precipitation data for Western Europe from 1980 to 2010 for trend analysis"), the system returns a ranked list of datasets $(d_1, d_2, \dots)$ intended to reflect the user's notion of relevance. We focus on the ranking stage after standard access controls and faceted filters have determined an eligible candidate set; RC-Rank is designed to be compatible with existing metadata-centric retrieval stacks that first retrieve a candidate pool and then apply a more expressive ranking function.

In practice, many portals implement ranking functions of the form

$$s(q, d) = \mathrm{BM25}(q, \mathrm{metadata}(d)) + \lambda \cdot \mathrm{recency}(d) + \cdots,$$

where a lexical similarity measure such as BM25, computed over a small set of textual fields (typically title, abstract, and keywords), is combined with one or two global signals such as publication date or download count. Variants of this pattern appear in many open data portals and institutional repositories and are often inherited from generic search platforms originally designed for document retrieval rather than data retrieval [1]. Such designs implicitly treat relevance as primarily topical, with weak secondary adjustments for timeliness or popularity, and they tend to combine heterogeneous signals without a clear semantic interpretation of what each term represents.

However, studies of scientific data relevance show that users jointly consider at least five interdependent criteria when deciding whether to reuse a dataset. Topicality captures the semantic match between the dataset content and the information need. Authority reflects trust in the data producer, curator, or hosting platform and is often inferred from institutional reputation, curated status, and reuse in the literature. Quality refers to perceived methodological soundness, completeness, and documentation, including the presence of clear methods, error estimates, and quality flags. Accessibility concerns licensing, cost, technical access mechanisms, and file formats, which determine whether users can legally and practically obtain and process the data. Finally, utility captures the task- and context-dependent suitability of the dataset for the user's intended analysis, including whether its granularity, coverage, and structure align with the requirements of a specific workflow [5, 6]. These criteria are not evaluated in isolation: empirical evidence suggests that users often begin with topicality, proceed to reliability judgments that integrate authority, quality, and accessibility, and then assess utility in their particular situation [5].

The core problem, therefore, is to design a ranking framework that moves beyond monolithic lexical scoring and provides a principled, auditable way to incorporate structured feature sets corresponding to these empirically grounded criteria. Such a framework should (i) expose a transparent aggregation mechanism so that repository managers can understand and, where necessary, adjust the

influence of different criteria, (ii) support data-driven learning from user interactions while remaining robust to sparse or biased feedback, and (iii) be practical for deployment under the latency and feature-availability constraints of real portals. RC-Rank addresses this problem by defining criterion-specific scores that are explicitly normalized and comparable, and by learning or tuning criterion weights that govern the trade-offs among topicality, authority, quality, accessibility, and utility.

2. Background and Related Work

Dataset search has emerged over the last decade as a distinct research area concerned with matching user data needs to collections of datasets in repositories, portals, and across the web, rather than merely retrieving documents that describe those datasets. Chapman et al. characterise dataset search as an ecosystem of systems and practices that spans institutional and disciplinary repositories, government open data portals, research infrastructures, and general-purpose services such as Google Dataset Search [1]. Large-scale production systems such as Google Dataset Search further highlight the heterogeneity of sources and practices involved in exposing structured data on the web and the need for schema-agnostic indexing and ranking strategies [11]. Log analyses of dataset search services and data portals show that users arrive with varying levels of domain expertise, often search across multiple portals in a single task, and combine search, browsing, and social strategies to locate data that can be reused in their work [12, 13].

Most operational systems in this space adopt a common architectural pattern. They index metadata about datasets—including titles, descriptions, keywords, publishers, formats, and licenses—and expose keyword-based search with faceted filtering over a small set of structured attributes such as topic category, temporal coverage, or publisher. This architecture is straightforward and scalable, because it treats dataset records much like short documents and can therefore reuse decades of development in text-based information retrieval engines. However, observational and log-based studies have shown that metadata-centric designs face persistent limitations. Metadata is often incomplete, inconsistent across repositories, or expressed in language that does not align with how users articulate their data needs, making it difficult to assess what a dataset contains or whether it is analytically useful [3, 4, 12, 14, 16]. As a result, ranking based solely on textual similarity can fail to surface datasets that are fit for a user's task but described with different vocabulary, while highly visible records may be topically related yet methodologically unsuitable. Empirical work on how researchers search for data also shows that they frequently rely on contextual cues and external documentation to infer fitness for use, which are only weakly reflected in standard metadata fields [14–16].

Scientific data portals, for example in climate science, genomics, or social science, often adopt the same basic search architecture as general-purpose open data portals but operate under different constraints. They may maintain richer, domain-specific schemas, controlled vocabularies, and curation workflows, and they can expose additional services such as subsetting, visualisation, or programmatic access. Nevertheless, in many of these systems the ordering of search results remains largely metadata- and keyword-centric, with only modest use of richer signals such as provenance, methodological descriptors, and usage traces [1, 12, 13]. Studies of user interaction with these portals suggest that scientists engage in complex sensemaking beyond typing a few keywords and clicking the first result: they inspect documentation, compare multiple candidate datasets, follow cross-links to related publications, and draw on disciplinary knowledge and social networks to assess what a dataset can and cannot be used for [3, 6, 14, 15]. Yet the ranking functions that determine which

datasets are shown first rarely encode these richer aspects of quality, trust, and reuse history in a systematic way.

Information retrieval research has long established that relevance is multi-dimensional, involving topical, cognitive, situational, and affective components that depend on properties of the information objects as well as the user's goals and context [2]. For scientific data, this multi-dimensionality is especially pronounced, because reusing a dataset entails non-trivial investment in understanding the methods by which the data were produced, the structure and semantics of the variables, and the limitations imposed by sampling, measurement error, and processing steps. Unlike document retrieval, where reading a single article may suffice to answer an immediate question, data reuse often involves integrating a dataset into an analysis workflow, and the consequences of misjudging relevance can be substantial.

Empirical work focused specifically on scientific data relevance has converged on a small set of core criteria that users repeatedly invoke when judging whether a dataset is suitable for reuse. Liu et al. identify five criteria as particularly salient: topicality, authority, quality, accessibility, and usefulness or utility [5]. Topicality concerns semantic match and is typically assessed through titles, abstracts, and keywords. Authority reflects trust in the producer, curator, or hosting platform and is inferred from institutional reputation, community endorsement, or inclusion in curated collections. Quality concerns methodological soundness, completeness, and documentation, including the clarity of methods descriptions, the availability of error estimates, and the presence of explicit quality flags. Accessibility refers to licensing, cost, access mechanisms, and formats that determine whether users can legally and practically obtain and process the data. Utility captures task- and context-dependent suitability, including whether temporal and spatial coverage, granularity, and variable selection match the requirements of a specific workflow. Related studies of data summarisation and data descriptions likewise show that users look for concise yet rich representations that help them quickly judge these criteria without downloading and inspecting the full data [16].

Importantly, these criteria are not applied as a static checklist. Liu et al. show that users tend to follow a process-oriented pattern in which they first make an initial topicality judgment to filter out manifestly irrelevant datasets, then engage in reliability judgments that combine authority, quality, and accessibility signals, and finally evaluate utility in relation to the specific task they have in mind [5]. Other qualitative studies of data search and data sensemaking report similar dynamics, with users moving iteratively between understanding "what the dataset is about", "whether it can be trusted", and "what it can be used for" [4, 6, 14, 15]. These studies also highlight that relevance judgments are shaped by social and organisational context: datasets produced by a user's own institution or by a well-known consortium may be granted higher initial authority, while institutional policies or disciplinary norms may constrain which licenses and access conditions are acceptable.

Despite this growing body of work, much prior research on user relevance criteria for scientific data has focused on understanding cognition and modelling paths between criteria at a conceptual level. Structural equation models capture how topicality, quality, authority, and accessibility jointly influence perceived usefulness [5], and qualitative frameworks describe the narratives and heuristics that data users employ when navigating complex portals [6, 14, 15]. Log analyses and session-based studies complement this picture by characterising how queries, clicks, and reformulations unfold over time [8, 9]. However, comparatively few studies translate these insights into concrete ranking algorithms or demonstrate, via controlled evaluation, how incorporating user-derived criteria into ranking functions changes retrieval effectiveness or user experience. There remains a gap between

knowing which criteria matter and having operational methods to exploit them at scale in real systems.

In web search, recommender systems, and related information access applications, it is widely recognised that no single signal suffices to capture relevance. Modern ranking functions combine heterogeneous evidence, including content similarity, link structure, freshness, usage statistics, and user preferences, which has led to an extensive literature on multi-factor ranking and learning-to-rank [10, 17, 18]. In learning-to-rank, feature vectors representing query–document pairs are mapped to scores by models trained on observed preferences or relevance judgments, with formulations that directly target ranking-quality metrics such as NDCG [17]. Within this broader context, there have also been explicit attempts to frame ranking as a multi-criteria decision problem in which different classes of features are treated as criteria with preference weights. Wolfe and Zhang investigate user-centric multi-criteria information retrieval where signals such as recency, cost, and authority are combined with topical relevance [19]. Attia et al. propose a multi-criteria indexing and ranking model that groups features into semantic, structural, and usage-related categories and combines them using multi-criteria decision-making techniques [10]. Their results, together with work in multi-objective ranking and fairness-aware re-ranking, suggest that multi-criteria approaches can yield more relevant and trustworthy rankings than single-criterion baselines, especially when users care about attributes beyond topical match [19–21].

However, despite the conceptual compatibility between multi-criteria ranking methods and the multi-dimensional nature of scientific data relevance, there has been limited exploration of multi-criteria ranking tailored specifically to datasets where the criteria themselves are derived from empirical models of data relevance. Existing multi-criteria models typically assume generic criteria such as content similarity, authority, and freshness and may not capture the nuances of data quality, accessibility, and task-specific utility that are central to scientific data search. Moreover, they rarely make explicit the mapping between cognitive constructs (such as users' notions of trust or usefulness) and concrete system features. This creates an opportunity to design ranking frameworks that leverage the methodological machinery of multi-criteria ranking and learning-to-rank while grounding the criteria and the interpretation of trade-offs in empirically validated accounts of how researchers judge datasets.

3. Problem Formulation and the RC-Rank Framework

In order to translate empirically observed relevance criteria into a concrete ranking mechanism, we formalize the scientific data search setting and introduce notation that makes the assumptions and design degrees of freedom explicit. We then describe the criterion-aware architecture of RC-Rank, outline criterion-specific feature instantiations that are implementable in typical portals, and specify how criterion scoring functions and aggregation weights can be configured or learned while maintaining interpretability.

3.1. Formal setting

Let $\mathcal{D}$ denote a collection of datasets indexed by a scientific data portal. Each dataset $d \in \mathcal{D}$ is assumed to be described by a metadata record M_d that contains fields such as title, description, subject keywords, publisher, license, format, temporal and spatial coverage, and any domain-specific attributes supported by the repository schema. In addition to this structured metadata, the portal

may maintain a set of derived statistics S_d for each dataset, capturing aspects such as basic data profiling (for example, missingness rates, value distributions, or the presence of obvious outliers), usage and impact (for example, download counts, citation counts, or mentions in data papers), and provenance or curation status. In practice, both M_d and S_d are heterogeneous across repositories. RC-Rank therefore assumes only that a portal can expose a minimal common subset and that missing values are handled explicitly (e.g., by conservative defaults and missingness indicators) rather than silently treated as zeros, because missingness itself is informative in metadata-centric settings.

Let $\mathcal{Q}$ denote the space of user information needs expressed as queries. A query $q \in \mathcal{Q}$ may consist of a short keyword string, a longer free-text description, or a more explicit task formulation such as "daily precipitation data for Europe from 1980 to 2010 for trend analysis". The portal implements a ranking function

$$s : \mathcal{Q} \times \mathcal{D} \to \mathbb{R},$$

where $s(q, d)$ measures the estimated relevance of dataset d to query q. For a given query, the system returns datasets ordered in decreasing values of $s(q, d)$, typically subject to additional constraints such as access permissions or user-configured filters. Conceptually, we view ranking as operating over an eligible candidate set determined by the retrieval stage and by filters (e.g., access constraints, temporal facets, or domain facets). RC-Rank is compatible with both single-stage ranking and two-stage pipelines that first retrieve a candidate pool and then apply a criterion-aware re-ranking step; the framework concerns how to compute and combine relevance signals within the ranking function.

In many existing systems, $s(q, d)$ is implemented as a lexical similarity score between q and a concatenation of textual metadata fields, often based on BM25 or a related probabilistic retrieval model, optionally combined with one or two simple global adjustments for recency or popularity. Formally, this can be written as

$$s_{\text{lex}}(q, d) = \text{BM25}(q, \text{metadata}(d)) + \lambda \cdot \text{recency}(d) + \mu \cdot \text{popularity}(d),$$

where λ and μ are hand-tuned constants and $\text{metadata}(d)$ denotes the subset of M_d used for indexing. Such a formulation effectively treats relevance as a function of topical relatedness, with weak secondary influence from a small number of non-topical signals. It does not differentiate between different classes of non-topical evidence, nor does it provide a natural mechanism for incorporating user-derived relevance criteria beyond ad hoc feature engineering.

RC-Rank starts from the observation that users judge the relevance of scientific datasets along several distinct dimensions, including topicality, authority, quality, accessibility, and task-specific utility. To capture this, we introduce a set of relevance criteria $\mathcal{C} = \{\text{T}, \text{A}, \text{Q}, \text{C}, \text{U}\}$, where T denotes topicality, A denotes authority, Q denotes quality, C denotes accessibility, and U denotes utility. Here $\mathcal{C}$ denotes the set of criteria, while the symbol C within the set labels the accessibility criterion; we retain this notation to keep criterion labels compact. For each criterion $c \in \mathcal{C}$, we define a criterion-specific score

$$f_c(q, d) \in [0, 1],$$

which aggregates a subset of features relevant to that criterion into a normalized value on the unit interval. The overall ranking score is then expressed as a weighted combination of these criterion scores,

$$s(q, d) = \sum_{c \in \mathcal{C}} w_c \, f_c(q, d),$$

where $w_c \geq 0$ are criterion weights that satisfy $\sum_{c \in \mathcal{C}} w_c = 1$. The functions $f_c(q, d)$ and the weights w_c are the primary objects of design and learning in RC-Rank. The key modelling assumption is that the trade-offs among criteria can be represented as a convex combination of criterion scores on a comparable scale; this enables interpretability and policy control, while still allowing each f_c to be implemented as a non-linear mapping of criterion-specific features.

A practical implication of this formulation is that score calibration matters. Because the weights w_c are meaningful only when the $f_c(q, d)$ values are comparable across criteria, RC-Rank requires each f_c to be normalized to the same range and to be robust to outliers and missing metadata. In our setting, "normalized to $[0, 1]$" means that criterion scores should be monotonically increasing with respect to the underlying criterion evidence and that the mapping should be stable across queries. Portals can operationalize this in several ways, including per-query normalization over the candidate set (e.g., robust min–max after clipping extreme values), global calibration using held-out interaction data, or a hybrid approach in which raw scores are globally calibrated and then lightly normalized within each query to control for query-to-query scale shifts. The remainder of this section specifies feature blocks and scoring functions in a way that is compatible with these calibration requirements.

3.2. Criterion-aware architecture

The RC-Rank framework can be viewed as a modular architecture layered on top of the formal setting described above. Given a query q and a dataset d, the system constructs a feature vector

$$\phi(q, d) = (\phi_{\mathrm{T}}(q, d), \phi_{\mathrm{A}}(d), \phi_{\mathrm{Q}}(d), \phi_{\mathrm{C}}(d), \phi_{\mathrm{U}}(q, d)),$$

where $\phi_{\mathrm{T}}(q, d) \in \mathbb{R}^{k_{\mathrm{T}}}$ denotes the block of topicality features, $\phi_{\mathrm{A}}(d) \in \mathbb{R}^{k_{\mathrm{A}}}$ denotes the block of authority features, $\phi_{\mathrm{Q}}(d) \in \mathbb{R}^{k_{\mathrm{Q}}}$ denotes the quality features, $\phi_{\mathrm{C}}(d) \in \mathbb{R}^{k_{\mathrm{C}}}$ denotes the accessibility features, and $\phi_{\mathrm{U}}(q, d) \in \mathbb{R}^{k_{\mathrm{U}}}$ denotes the utility features. Some of these blocks, such as topicality and utility, depend explicitly on both q and d, while others, such as authority, quality, and accessibility, depend only on the dataset and its context. This separation is intentional: it makes explicit which parts of the ranking function are query-sensitive and which are dataset priors that should not fluctuate wildly across queries.

Each block of features is transformed into a scalar criterion score via a criterion-specific scoring function

$$f_c(q, d) = g_c(\phi_c(q, d); \theta_c),$$

where g_c denotes a parametric mapping and θ_c denotes parameters associated with criterion c. In deployments, g_c can range from a linear model with bounded output (e.g., a linear score passed through a sigmoid and then calibrated to $[0, 1]$) to a shallow non-linear model such as a small gradient-boosted tree ensemble trained on criterion-relevant features only. The design requirement is that g_c should be stable under missing values, should allow monotonic behaviour where appropriate (e.g., higher metadata completeness should not reduce the quality score), and should output scores that are comparable across criteria after normalization.

The outputs of these scoring functions are combined linearly with weights w_c to yield the final ranking score $s(q, d)$. This explicit aggregation layer separates two sources of modelling choice: (i) how criterion evidence is extracted and mapped into f_c via ϕ_c and g_c, and (ii) how the portal wishes to trade off criteria in a given community via w_c. The decomposition is also the basis for criterion-level explanation: at ranking time, the system can report the vector $(f_{\mathrm{T}}(q, d), f_{\mathrm{A}}(q, d), f_{\mathrm{Q}}(q, d), f_{\mathrm{C}}(q, d), f_{\mathrm{U}}(q, d))$ and the corresponding weighted contributions $w_c f_c(q, d)$.

Figure 1 conceptually illustrates this architecture, with separate channels for each criterion feeding into an aggregation layer.

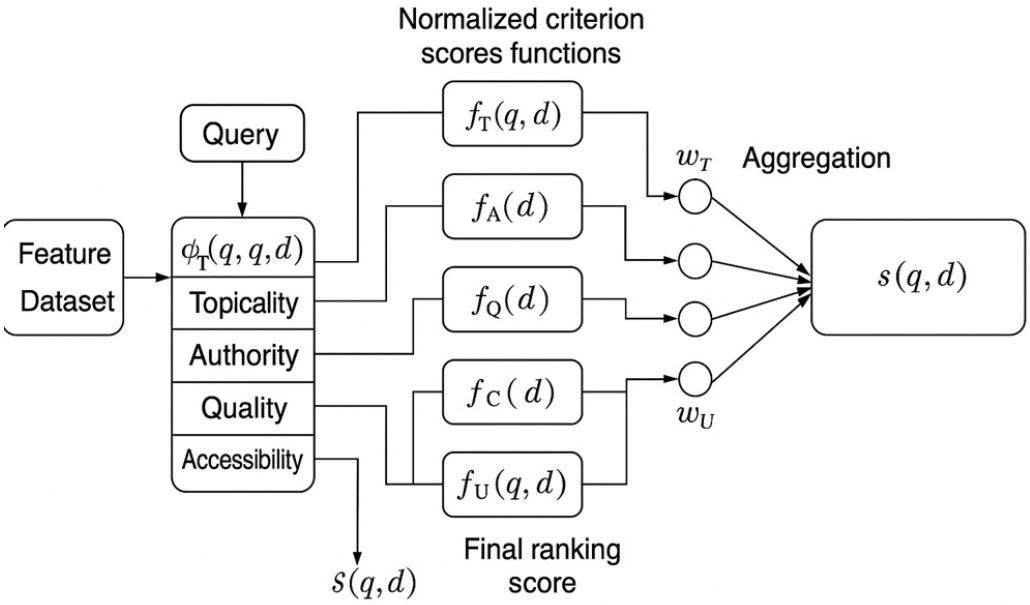

Fig. 1. Conceptual architecture of RC-Rank. Features are grouped into criterion-specific channels, transformed into normalized criterion scores $f_c(q,d)$, and aggregated by weights w_c into the final ranking score $s(q,d)$

This modular design has several advantages. It supports simple linear scoring within each criterion when data are scarce, as well as more complex non-linear transformations when sufficient interaction data are available. It also allows designers to inspect and debug each criterion in isolation, to adjust weights based on policy or domain knowledge, and to communicate ranking behaviour in criterion terms rather than as an opaque mixture of unrelated signals.

3.3. Criterion-specific features

RC-Rank remains agnostic about the exact choice of features, but it assumes that each criterion can be instantiated using information typically available in scientific data portals. In what follows, we specify concrete, reproducible feature designs for each criterion and clarify the assumptions required to compute them in heterogeneous portals.

Topicality captures semantic match between the dataset and the information need expressed by q. In most portals, this can be approximated using textual metadata fields such as title, description, subject keywords, and domain tags. A topicality feature block can include fielded BM25 similarities between q and each textual field (with standard BM25 parameters as a default and field weights tuned on held-out queries), query-term overlap with controlled keywords, and an embedding-based cosine similarity between vector representations of the query and the concatenated textual metadata. Because these signals can have different raw scales, they should be normalized in a way that preserves ordering within each query (e.g., by scaling each component to $[0,1]$ over the candidate set before combining). The topicality criterion score $f_T(q,d)$ can then be defined as a weighted combination of these normalized components followed by a final calibration step to ensure that $f_T(q,d)$ is comparable to other criterion scores.

Authority reflects trust in the dataset's producer, curator, and hosting context. In scientific data portals, relevant signals include the type and reputation class of the publisher or institution,

membership in curated or flagship collections, and usage and impact proxies such as download and citation counts. To make usage signals comparable across dataset ages, counts should be transformed (e.g., using $\log(1+x)$) and normalized by dataset age or exposure time, and categorical provenance indicators (e.g., "curated", "peer-reviewed data paper available", "official statistics") should be encoded explicitly rather than inferred indirectly. A simple authority scoring function $f_{\mathrm{A}}(q,d)$ can compute a bounded combination of these features and normalize the result, with the understanding that f_{A} is query-independent but still participates in the final ranking.

Quality refers to methodological soundness, completeness, and internal consistency of a dataset and its documentation. Although some aspects of data quality are domain-specific, portals increasingly record generic indicators such as metadata completeness ratios, the presence and length of methodological descriptions, links to quality reports or validation studies, and basic profiling statistics such as fraction missing, existence of quality flags, or obvious format anomalies. These signals can be aggregated into a quality feature block $\phi_{\mathrm{Q}}(d)$ that includes (i) a completeness ratio defined over a fixed field inventory, (ii) documentation richness indicators (e.g., presence of a methodology section, codebook, or data dictionary), and (iii) lightweight profiling features where available. Because quality evidence is often missing for legacy datasets, the scoring function for f_{Q} should treat missingness explicitly and remain conservative (e.g., missing documentation should not inflate the score). The quality score is query-independent but operationally important because it captures the reliability dimension that users repeatedly emphasize when moving from topical filtering to reuse decisions.

Accessibility encompasses licensing, cost, technical access mechanisms, and file formats. From metadata, this includes the license type (e.g., permissive versus restricted), presence of persistent identifiers, availability of machine-readable formats (such as CSV or NetCDF) as opposed to image- or PDF-only representations, and the existence of APIs or bulk download options. These factors can be encoded as ordinal features (for openness of license and access conditions) and as binary or graded indicators (for machine-readability and access modalities). For very large datasets, approximate indicators of size or typical access latency can also be included, because extreme size can be a practical barrier for some users even when access is legally open. The accessibility criterion score $f_{\mathrm{C}}(d)$ consolidates these features into a single measure of how easy it is, in principle, to obtain and reuse the dataset under common research workflows.

Utility (usefulness) is the most task- and user-dependent criterion. It captures whether a dataset that is topical, authoritative, high-quality, and accessible is suitable for the specific analysis a user intends to perform. Utility features therefore combine metadata, query-derived constraints, and interaction evidence. They can include explicit constraint matches inferred from the query text (e.g., temporal range, spatial coverage, resolution, population, or unit of analysis), similarity between the query and documented use cases or associated publications, and reuse signals indicating that a dataset has supported tasks similar to the current query (e.g., trend analysis, model calibration, or benchmarking). Because many of these aspects are difficult to infer purely from static metadata, $f_{\mathrm{U}}(q,d)$ is a natural target for learning from implicit feedback such as clicks, downloads, dwell times, and session-level satisfaction proxies. Importantly, the criterion-aware structure allows utility modelling to become more sophisticated over time without collapsing interpretability, because the utility channel remains distinct from topicality and reliability channels.

3.4. Aggregation and learning

Given criterion scores $f_c(q, d) \in [0, 1]$ for all $c \in \mathcal{C}$, RC-Rank combines them into a single ranking score according to

$$s(q, d) = \sum_{c \in \mathcal{C}} w_c \, f_c(q, d),$$

with non-negative weights w_c that sum to one. This aggregation rule admits both expert-driven and data-driven instantiations. In early deployment scenarios, domain experts such as data librarians, repository managers, or experienced users can specify initial weights that reflect typical user priorities; for example, topicality may receive the largest share, reliability-related criteria (authority and quality) may jointly receive a substantial share, and accessibility and utility may receive the remainder. This yields an interpretable baseline that can be inspected and communicated to stakeholders.

As interaction data accumulate, the same formulation supports learning the weights w_c from user behavior while respecting the simplex constraint. One practical approach is to treat the criterion scores as features in a constrained linear model and estimate w_c by minimizing a ranking loss (e.g., a pairwise loss based on observed preferences) with non-negativity and sum-to-one enforced via projection onto the simplex after each optimization step. A complementary approach is to learn the criterion mappings $g_c(\cdot; \theta_c)$ and the weights jointly, but with regularization that shrinks parameters toward conservative priors to avoid overfitting in low-traffic settings. In high-data regimes, it is possible to learn user- or group-specific weights, yielding role-specific ranking behaviour in which, for example, teaching-focused users receive rankings that emphasize accessibility and pedagogical utility, while regulatory analysts see rankings that emphasize authority and methodological quality.

RC-Rank also supports learning at multiple levels. At the feature-to-criterion level, parameters θ_c can be learned from explicit judgments or from implicit feedback using probabilistic models that predict observed engagement conditional on q and d. At the criterion level, the weights w_c can be adapted over time, potentially with temporal smoothing to account for drift in community priorities. In settings where interaction data are sparse or unevenly distributed across topics, a hybrid strategy is attractive: start from hand-tuned weights and simple, monotonic scoring functions, monitor stability and diagnostic signals (e.g., distribution of criterion scores and the frequency of missing evidence), and then gradually refine both g_c and w_c as more data accumulate. The criterion-aware structure of RC-Rank provides a scaffold for these learning processes and makes it easier to reason about how model changes relate back to the underlying constructs of topicality, authority, quality, accessibility, and utility.

4. Illustrative Case Study

To illustrate how RC-Rank behaves once instantiated with criterion scores and weights, consider a national agricultural data portal that aggregates statistics from government agencies, research institutes, and regional experimental stations. A researcher interested in food security issues the query

> "Wheat yield data for China from 2010 to 2015 for trend analysis."

The portal indexes a large number of agricultural datasets, many of which mention wheat or crop yields in some form. After applying basic access and domain filters, suppose that three candidate datasets are particularly plausible for this query. The case study is illustrative: the datasets

and criterion scores are described at a level of detail typical of portal metadata, and the numeric criterion scores are intended to be plausible outputs of the scoring functions rather than empirical measurements.

The first dataset, denoted D_1, is a national crop yield dataset produced by a provincial statistics office and later harvested into the national portal. It covers multiple crops and years at the provincial level, and its description explicitly mentions wheat yields for the period 2005–2018. The metadata provide a concise title and a short description, but methodological documentation is limited to a few sentences and there is no separate quality report. The publisher is a mid-level regional agency that is not widely known outside national circles, and the dataset has been downloaded relatively few times since publication.

The second dataset, D_2, is a curated agricultural statistics product published by the national bureau of statistics. It aggregates provincial reports into a harmonised time series for major crops, including wheat, and provides national and provincial yields from 1990 onward. The title and high-level description focus on agriculture rather than wheat specifically, so the textual match to the query is slightly less direct. However, D_2 has extensive methodological documentation, including descriptions of sampling and estimation procedures, revision policies, and links to official statistical yearbooks. It is hosted on a flagship government portal under an open licence that explicitly permits reuse for research, and it has a long history of downloads and citations in agronomy and economics journals.

The third dataset, D_3, is a regional agronomic experiment dataset produced by a leading agricultural university. It consists of plot-level measurements from fertiliser and irrigation experiments conducted in a specific wheat-growing region between 2012 and 2014. The dataset includes detailed variable descriptions, measurement protocols, and calibration information, and a companion data paper describes the experimental design in depth. Access, however, is restricted: users must request permission, and the licence limits reuse to non-commercial scientific purposes with additional constraints. The geographical coverage is narrower than the national scope implied in the query, but the granularity is much finer.

Assume that the portal has computed normalized criterion scores for these three datasets on the $[0, 1]$ scale based on the feature designs outlined in Section 3. Table 1 presents one plausible configuration of scores for topicality, authority, quality, accessibility, and utility.

Dataset	Topicality	Authority	Quality	Accessibility	Utility
D_1	0.90	0.40	0.60	0.80	0.50
D_2	0.80	0.90	0.80	0.70	0.70
D_3	0.70	0.60	0.90	0.50	0.60

Table 1. Normalized criterion scores for three agricultural datasets in the wheat-yield case study.

In this configuration, D_1 achieves the highest topicality score because its title and description mention wheat yields and the 2010–2015 period explicitly. Its authority score is moderate to low due to the lesser-known publisher and limited usage, and its quality score reflects the presence of basic metadata but minimal methodological detail. Accessibility is relatively high because the data are available as downloadable CSV files under a permissive licence. Utility is moderate: the dataset covers the requested time period and geographic scope, but the lack of detailed documentation may make it harder to use in rigorous trend analyses.

Dataset D_2 has slightly lower topicality in textual terms, because the title emphasises general agricultural statistics rather than wheat specifically. However, its authority and quality scores are high, reflecting the reputation of the national bureau of statistics, the extensive methodological documentation, and the widespread reuse and citation of the dataset. Accessibility is somewhat lower than for D_1 due to less convenient file formats and the need to access some components through a bulk download interface, but the open licence keeps it reasonably high. Utility is also high: the dataset provides consistent, long-term series that are well suited for national-level trend analysis.

Dataset D_3 has the lowest topicality score of the three, because its title and abstract emphasise regional field experiments and agronomic treatments rather than national yields. Its authority score is moderate: the producing institution is a respected university, but the dataset is less visible in national policy circles. The quality score is very high, reflecting the rich documentation and detailed experimental protocols. Accessibility is constrained by the restrictive licence and the need for permission, and the narrower spatial coverage means that its utility for a national trend analysis is moderate rather than high, even though it may be extremely valuable for other tasks such as modelling yield response to fertiliser.

Using hand-tuned criterion weights that reflect a generic research-oriented search scenario,

$$w_{\mathrm{T}} = 0.35, \quad w_{\mathrm{A}} = 0.20, \quad w_{\mathrm{Q}} = 0.20, \quad w_{\mathrm{C}} = 0.15, \quad w_{\mathrm{U}} = 0.10,$$

we obtain the overall RC-Rank scores

$$\begin{aligned} s(q, D_1) &= 0.35 \cdot 0.90 + 0.20 \cdot 0.40 + 0.20 \cdot 0.60 + 0.15 \cdot 0.80 + 0.10 \cdot 0.50 \\ &= 0.685, \\ s(q, D_2) &= 0.35 \cdot 0.80 + 0.20 \cdot 0.90 + 0.20 \cdot 0.80 + 0.15 \cdot 0.70 + 0.10 \cdot 0.70 \\ &= 0.795, \\ s(q, D_3) &= 0.35 \cdot 0.70 + 0.20 \cdot 0.60 + 0.20 \cdot 0.90 + 0.15 \cdot 0.50 + 0.10 \cdot 0.60 \\ &= 0.680. \end{aligned}$$

RC-Rank therefore orders the datasets as $D_2 \succ D_1 \succ D_3$. The curated government dataset D_2 is ranked highest, even though D_1 has a slightly better topicality score, because D_2 has substantially higher authority, quality, and utility. This ordering is consistent with the qualitative decision pattern described in prior work: once topicality is adequate, users often privilege reliability-related signals (authority and quality) and then consider task fit (utility) before committing to reuse. At the same time, the ordering is contingent on the explicit weight configuration, which makes the induced trade-off inspectable.

A useful robustness check in the criterion-aware setting is sensitivity to plausible changes in weights. If topical match dominates because the user is browsing broadly and is willing to tolerate weaker documentation, one can increase the topicality weight substantially, for example to $w_{\mathrm{T}} = 0.75$ with the remaining weight distributed evenly across the other criteria ($w_{\mathrm{A}} = w_{\mathrm{Q}} = w_{\mathrm{C}} = w_{\mathrm{U}} = 0.05$). Under this configuration, the resulting scores become $s(q, D_1) = 0.79$, $s(q, D_2) = 0.755$, and $s(q, D_3) = 0.655$, which changes the ordering to $D_1 \succ D_2 \succ D_3$. Conversely, if the task is high-stakes and the portal wishes to be conservative with respect to trust and documentation, shifting weight from topicality toward authority and quality further separates D_2 from less documented sources. These simple perturbations illustrate two properties that matter operationally: (i) the ranking changes in

predictable ways when weights change, and (ii) the criterion contributions provide a direct explanation of why the ordering changes.

Finally, note that the numeric criterion scores are illustrative and are used here to demonstrate the mechanics and interpretability of RC-Rank. The empirical question of whether criterion-aware ranking improves retrieval effectiveness and user satisfaction depends on how criterion scores are computed from real metadata and interaction signals and must therefore be answered through the evaluation protocol in the evaluation protocol section below.

5. Evaluation Protocol

To move from conceptual design to empirical validation, RC-Rank must be evaluated in realistic settings and compared against ranking approaches already used in scientific data portals. Because dataset search decisions are influenced by both ranking effectiveness and by users' trust and sense-making practices, an adequate evaluation programme should combine offline experiments based on historical interaction data with prospective user studies that capture perceived relevance, trust, and usability.

A natural starting point is to define a small set of research questions that guide the evaluation. One central question is whether RC-Rank improves ranking effectiveness compared to keyword-based baselines when measured by standard information retrieval metrics such as normalized discounted cumulative gain (NDCG) and mean average precision (MAP) on held-out queries. A second question is whether users perceive rankings produced by RC-Rank as better aligned with their own relevance judgments and decision processes than those produced by existing systems. A third question concerns sensitivity: to what extent do retrieval outcomes depend on the choice of criterion weights and on the feature instantiations used to compute each $f_c(q, d)$? Finally, it is important to understand whether learned criterion weights are stable over time and whether they match qualitative expectations derived from prior relevance studies and expert intuition.

For offline learning-to-rank evaluation, we assume access to a log of search interactions in a scientific data portal. The log should contain representative queries or search tasks, the ranked lists that were presented, and explicit or implicit relevance signals for query–dataset pairs (for example, clicks, downloads, bookmarks, outbound API calls, or other meaningful engagement). Because click and download logs are biased by result position and by interface design, the evaluation methodology must explicitly address these biases rather than treating raw clicks as ground-truth relevance. A practical approach is to construct training labels using multiple signals (e.g., treating a download or bookmark as stronger evidence than a click) and to incorporate debiasing via inverse propensity weighting or a standard click model that estimates examination probability as a function of rank position. When such modelling is not feasible, a conservative alternative is to restrict labels to high-intent actions (e.g., downloads) and to treat unexamined results as missing rather than negative.

From this log, a query set should be constructed that captures the diversity of information needs encountered in the portal. To avoid temporal leakage, queries can be split chronologically into training, validation, and test periods, and repeated queries by the same users should be handled carefully to prevent memorization effects. For each query, a candidate set should be generated using the portal's production retriever (typically a fielded BM25 over metadata) to a fixed depth K (e.g., top-100 or top-1000), and then all models should re-rank the same candidate pool to ensure a fair comparison. For each query–dataset pair in the candidate set, the feature vector $\phi(q, d)$ and criterion

scores $f_c(q, d)$ are computed using the feature designs described earlier, with explicit handling of missing metadata (e.g., missingness indicators and conservative defaults) so that missing evidence does not spuriously inflate scores.

Baselines for comparison should be chosen to reflect what portals realistically deploy. A first baseline is a purely lexical model such as BM25 applied to selected metadata fields, with the BM25 parameters and field weights either set to standard defaults or tuned on a validation set. A second baseline is an enhanced lexical model that adds simple non-topical adjustments for recency and popularity, for example by adding normalized versions of $\log(1 + \text{downloads})$ and an age-based decay term; crucially, these additional terms should be calibrated so that their magnitude is commensurate with the lexical score. Beyond these, it is informative to include a standard learning-to-rank baseline trained on the full feature vector $\phi(q, d)$ without criterion grouping, because this quantifies the value of the criterion-aware decomposition relative to a purely predictive model.

RC-Rank should be instantiated in multiple variants to test which design choices matter. At minimum, one variant should use hand-tuned criterion weights to reflect a reasonable default portal policy; another should learn the weights w_c over the criterion scores $f_c(q, d)$ under the non-negativity and sum-to-one constraints; and a third can learn a richer model over the full $\phi(q, d)$ while still reporting criterion-block contributions for interpretation. Hyperparameters, including regularization strength, learning rate, and any non-linear model settings, should be tuned using the validation set only, with the test set used once for final reporting. Performance should be reported using NDCG at cut-off positions such as 10 and 20, MAP at comparable cut-offs, and recall at broader depths, with per-query metric distributions reported in addition to means to show whether gains are driven by a small subset of queries or are consistent across the workload.

To strengthen rigor, the analysis should quantify uncertainty and robustness. Statistical uncertainty can be reported via paired bootstrap confidence intervals over queries for NDCG and MAP, and significance can be assessed using paired randomization tests or bootstrap-based hypothesis tests that respect the paired nature of per-query outcomes. Robustness checks should include: (i) sensitivity to the candidate pool depth K (to ensure that improvements are not an artefact of a particular truncation depth), (ii) sensitivity to reasonable perturbations of w_c around learned or hand-tuned values (to ensure that conclusions are not driven by a fragile weight configuration), and (iii) ablation studies in which each criterion channel is removed or replaced by a naive proxy to quantify its marginal contribution. Additional stress tests can target practical deployment concerns, such as performance on "cold-start" datasets with minimal usage history, performance under artificially masked metadata fields (to emulate incomplete records), and temporal generalization when models are trained on earlier logs and evaluated on later periods.

While offline metrics provide evidence about ranking quality relative to historical behaviour or explicit labels, they do not fully capture users' perceptions of relevance, trust, and transparency. For this reason, a complementary user study is needed. A realistic design is a within-subject experiment with domain experts and frequent data users recruited from the portal's community. Participants are presented with a series of search tasks that reflect real data needs, such as constructing a climate index, replicating a published analysis, or preparing a teaching dataset. For each task, they interact with two versions of the search interface: one backed by a baseline ranking and one backed by RC-Rank, with the order of conditions randomized to control for learning effects. After completing each task, participants provide relevance assessments for the top-ranked datasets and answer short questionnaires about trust, satisfaction, and perceived transparency of the ranking, while task outcome

measures such as time-to-first-usable dataset and the number of result inspections can be recorded unobtrusively.

During the study, think-aloud protocols and screen recordings can be used to trace the cues that participants attend to and to assess whether the ordering induced by RC-Rank aligns with the stepwise judgment patterns reported in prior research (from topical filtering to reliability assessment to utility assessment). Qualitative analysis of these narratives can reveal whether the criterion-aware design resonates with users' mental models and whether criterion-level explanations help or hinder decision making. Quantitative analysis should again report uncertainty, for example by using mixed-effects models or paired tests appropriate for within-subject designs and by reporting confidence intervals for task metrics.

Taken together, offline experiments and user studies provide a comprehensive evaluation protocol for RC-Rank. Offline metrics test whether encoding user-oriented criteria into the ranking function yields measurable gains in effectiveness under realistic biases and constraints, while user studies assess whether such gains translate into improved support for real-world data search practices. The criterion-aware structure of RC-Rank also facilitates diagnostic evaluation: by inspecting learned weights, per-criterion score distributions, and ablation outcomes, repository managers can identify which aspects of relevance the system captures well and where additional metadata, features, or interface support may be needed.

6. Discussion

RC-Rank can be seen as a bridge between two strands of work that have largely evolved in parallel: cognitive and empirical models of scientific data relevance on the one hand, and algorithmic designs for dataset search on the other. Research on relevance in information science has for decades emphasized that relevance is multi-dimensional and context-dependent, involving topical, cognitive, situational, and affective components [2]. More recent studies specific to research data refine this picture by identifying topicality, authority, quality, accessibility, and usefulness as salient criteria in scientific data users' decision processes [5, 6]. In contrast, many operational dataset search systems still deploy ranking functions that are essentially lexical, with limited incorporation of non-topical evidence and little explicit connection to these user-derived criteria [1]. RC-Rank addresses this gap by embedding empirically grounded relevance criteria into the structure of the ranking function, making the criteria explicit objects of system design rather than implicit by-products of feature engineering.

A central benefit of this design is interpretability that is directly tied to the criteria users articulate. Because RC-Rank decomposes the overall score $s(q, d)$ into criterion-specific components $f_c(q, d)$ with weights w_c, repository managers and system designers can inspect how much each criterion contributes to the ranking of specific datasets for specific queries. In the case study, for example, it becomes clear that a dataset can outrank a more textually matched alternative because stronger authority and quality evidence compensates for weaker topicality, which aligns with qualitative narratives in which trust and documentation become decisive once obviously irrelevant results are filtered out [3, 5]. Criterion-level inspection also supports debugging: if rankings appear counterintuitive, operators can determine whether the issue lies in topicality matching, in how provenance and quality are encoded, or in the chosen trade-off encoded by w_c.

Interpretability at the criterion level is also important from a governance perspective. Scientific data portals often operate under explicit policies or community norms concerning which kinds of

datasets should be promoted or downplayed. For instance, a repository may wish to ensure that official products are visible but not to the exclusion of high-quality research datasets, or it may want to guarantee that overly restrictive licensing does not dominate top ranks. In purely black-box learning-to-rank systems such policy goals are difficult to enforce or audit. In RC-Rank, by contrast, adjusting the weight associated with accessibility or authority has a clear and inspectable impact, and the resulting trade-offs can be discussed with stakeholders in terms of licensing, provenance, and documentation rather than opaque model parameters [21].

A second advantage of RC-Rank is flexibility and extensibility. Although the present formulation focuses on five criteria motivated by existing relevance studies, the framework can accommodate additional channels where justified by domain needs or policy constraints. Ethical compliance and privacy risk, for example, are increasingly salient in data-intensive domains where human subjects, sensitive environments, or regulated data are involved [22]. A future instantiation could include an explicit risk criterion grounded in metadata about consent, anonymization, and regulatory approvals, with its own feature block and weight. Similarly, representativeness or fairness may warrant explicit treatment in contexts where it matters whether datasets adequately cover different populations or regions and whether ranking choices create systematic exposure imbalances [23]. The architectural point is that adding such channels does not require changing the overall decomposition; it extends the set of criteria and preserves the interpretability and modularity that make the framework actionable.

Flexibility also extends to the choice of features and learning algorithms within each criterion. In low-resource settings where only basic metadata are available, $f_c(q, d)$ can be implemented using robust statistics such as metadata completeness ratios, fielded lexical similarities, and coarse-grained publisher indicators. As repositories evolve and richer information becomes available—for example, structured provenance records, standardized quality reports, or fine-grained usage analytics—these can be added incrementally as new features without disturbing the broader structure of RC-Rank. Likewise, when interaction data are sparse, simple monotonic mappings and hand-tuned weights may be appropriate; as more data accumulate, more expressive models can be used to learn θ_c and w_c while still reporting criterion-level summaries for interpretability [10, 24].

The explicit weight layer also supports principled personalization and task-aware ranking. In the current formulation, weights w_c are shared across users and queries, reflecting an assumed "average" scenario. However, different user groups and tasks prioritize criteria differently, and these differences can be represented explicitly via user- or group-specific weights learned from interaction patterns or configured through preferences. This opens the door to interfaces that allow users to adjust criteria such as trust, openness, or granularity and immediately see how rankings change, providing a controllable and transparent form of personalization rather than opaque behavioral targeting [21–23].

RC-Rank also clarifies the dependence of ranking quality on metadata and analytics. Many repositories struggle with incomplete or heterogeneous metadata, which directly limits what ranking algorithms can do. By making criterion scores explicit, the framework surfaces which metadata fields and derived statistics drive quality, accessibility, and utility assessments and therefore provides a concrete argument for investing in better documentation, provenance tracking, and usage monitoring. These incentives align naturally with broader FAIR goals, which emphasize not only findability and accessibility but also interoperability and reusability [8, 24].

At the same time, the framework has limitations that must be acknowledged to avoid over-claiming. First, criterion-aware aggregation does not eliminate the need for careful feature design and calibration: if a portal's authority or quality signals are weak, the corresponding criterion scores

may be noisy or biased. Second, because authority and usage-related signals can reinforce existing visibility advantages, naive use of such features may amplify feedback loops that favor well-established institutions and already popular datasets. Mitigating these risks may require explicit regularization of authority weights, stability constraints that prevent rapid swings driven by short-term popularity, and evaluation procedures that probe exposure and performance for new or under-documented datasets rather than only average metrics. Finally, the linear aggregation assumption is deliberately chosen for interpretability; if strong non-linear interactions among criteria are present in a given portal, empirical evaluation should test whether the interpretability benefits justify the potential loss of predictive flexibility.

7. Conclusion

We have proposed RC-Rank, a user-oriented multi-criteria ranking framework for scientific data search that explicitly incorporates empirically grounded relevance criteria into the ranking function. By structuring evidence into topicality, authority, quality, accessibility, and utility channels and aggregating them via interpretable weights, RC-Rank provides a practical mechanism for aligning ranking behavior with how researchers report judging dataset relevance.

We formalized the setting and design assumptions that make criterion-aware ranking operational in real portals, described concrete feature instantiations that can be derived from typical metadata and derived statistics, and clarified how criterion scoring functions and criterion weights can be configured or learned under constraints that preserve auditability. We also illustrated the framework on a synthetic example to show how different, explicitly stated weight configurations produce predictable and explainable ranking trade-offs, and we specified an evaluation protocol that supports rigorous offline and user-centered validation, including robustness checks and uncertainty reporting.

The primary contribution of this work is therefore methodological and architectural: it provides an explicit scaffold for building and evaluating criterion-aware dataset ranking systems. Demonstrating empirical gains on a deployed portal requires implementing the feature extraction and calibration pipeline on real metadata, learning from interaction logs while addressing bias, and conducting controlled evaluations with domain users. The evaluation protocol described here is intended to support that future empirical work and to ensure that any claims of effectiveness are directly supported by evidence.

References

[1] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L. D., Kacprzak, E., & Groth, P. (2020). Dataset search: a survey. *The VLDB Journal, 29*(1), 251-272.

[2] Saracevic, T. (2022). *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?*. Springer Nature.

[3] Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology, 70*(5), 419-432.

[4] Koesten, L., Gregory, K., Groth, P., & Simperl, E. (2021). Talking datasets–understanding data sensemaking behaviours. *International Journal of Human-Computer Studies, 146*, 102562.

[5] Liu, J., Wang, J., Zhou, G., Wang, M., & Shi, L. (2020). How do people make relevance judgment of scientific data?. *Data Science Journal, 19*, 9-9.

[6] Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science, 46*(4), 459-475.

[7] Rodríguez Rocha, O., & Faron Zucker, C. (2018, April). Automatic generation of quizzes from DBpedia according to educational standards. In *Companion Proceedings of the The Web Conference 2018* (pp. 1035-1041).

[8] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data, 3*(1), 1-9.

[9] Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., & Simperl, E. (2019). Characterising dataset search—An analysis of search logs and data requests. *Journal of Web Semantics, 55*, 37-55.

[10] Attia, M., Abdel-Fattah, M. A., & Khedr, A. E. (2022). A proposed multi criteria indexing and ranking model for documents and web pages on large scale data. *Journal of King Saud University-Computer and Information Sciences, 34*(10), 8702-8715.

[11] Brickley, D., Burgess, M., & Noy, N. (2019, May). Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference* (pp. 1365-1375).

[12] Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., & Simperl, E. (2019). Characterising dataset search—An analysis of search logs and data requests. *Journal of Web Semantics, 55*, 37-55.

[13] Carevic, Z., Roy, D., & Mayr, P. (2020, August). Characteristics of dataset retrieval sessions: experiences from a real-life digital library. In *International Conference on Theory and Practice of Digital Libraries* (pp. 185-193). Cham: Springer International Publishing.

[14] Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology, 70*(5), 419-432.

[15] Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science, 46*(4), 459-475.

[16] Koesten, L., Simperl, E., Blount, T., Kacprzak, E., & Tennison, J. (2020). Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies, 135*, 102367.

[17] Liu, T. Y. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval, 3*(3), 225-331.

[18] Qin, T., Liu, T. Y., Xu, J., & Li, H. (2010). LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval, 13*(4), 346-374.

[19] Wolfe, S. R., & Zhang, Y. (2009, July). User-centric multi-criteria information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 818-819).

[20] Gao, R., & Shah, C. (2020). Toward creating a fairer ranking in search engine results. *Information Processing & Management, 57*(1), 102138.

[21] Alvarez, P. A., Ishizaka, A., & Martinez, L. (2021). Multiple-criteria decision-making sorting methods: A survey. *Expert Systems with Applications, 183*, 115368.

[22] Oladipupo, O., Makpokpomi, O., & Adubi, S. (2023, April). A multi-criteria decision making approach to journal selection and ranking. In *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals* (SEB-SDG) (Vol. 1, pp. 1-11). IEEE.

[23] Szopik-Depczyńska, K., Cheba, K., & Wiśniewska, J. (2020). Innovation, R&D and user-driven innovation activity in R&D departments in Poland. The multi-criteria analysis. *Procedia Computer Science, 176*, 2705-2713.

[24] Liu, J., Wang, J., Zhou, G., & Zhang, G. (2023). *A Cognitive Model of Data-Quality Judgment in User Data Retrieval.* Available at SSRN 4465043.

How to cite this article: Mobeen Akhter and Nazish Noreen (2025). RC-Rank: A User-Oriented Multi-Criteria Ranking Framework for Scientific Data Search. *Bulletin of Computer and Data Sciences*, 6(1), 61-80. DOI: 10.71448/bcds2561-4

Received: 03/01/2025 **Revised:** 12/02/2024 **Accepted:** 26/02/2025 **Publish:** 30/03/2025