





# A Multi-component Similarity Measure for Personalized Content Discovery in Periodical Digital Library Collections

Emanuela Mitreva<sup>1</sup> , Desislava Paneva-Marinova<sup>1</sup> , Vladimir Georgiev<sup>2</sup> ,  
and Alexandra Nikolova<sup>1</sup> 

<sup>1</sup> Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS), Sofia,  
Bulgaria

emitreva@gmail.com, dessi@cc.bas.bg, alxnikolova@gmail.com,  
a.nikolova@math.bas.bg

<sup>2</sup> American University in Bulgaria, Blagoevgrad, Bulgaria  
vgeorgiev@aubg.edu

**Abstract.** The rapid growth of digital resources in modern digital libraries has intensified information overload, particularly in collections dominated by periodical publications. Such materials are inherently heterogeneous, multi-thematic, and often noisy, which complicates document modelling, similarity assessment, and the delivery of personalized content. The purpose of this research is to design and empirically validate a multi-component document similarity measure that supports robust and interpretable personalization in periodical-heavy digital libraries. Existing similarity approaches are largely designed for short or single-topic documents and, therefore, struggle to capture the partial and localized thematic overlap that characterizes periodicals. To address this gap, this study proposes a multi-component document similarity measure explicitly tailored to long, multi-topic periodical content. The proposed measure combines complementary perspectives on document relatedness by jointly accounting for global thematic orientation, localized content overlap, thematic distribution, and factual context within a unified and parameterizable formulation. By treating similarity as a composite phenomenon rather than a single-dimensional score, the approach enables more stable and meaningful identification of related documents in heterogeneous collections. The proposed measure is intended to support similarity-based navigation by helping users locate documents related to a specific document they are currently accessing or reviewing, thereby improving focused exploration within complex digital library collections.

**Keywords:** Multi-component similarity measure · Document similarity · Multi-topic documents · Digital libraries · Periodicals

## 1 Introduction

In recent decades, digital transformation led to fundamental changes in the way cultural heritage is stored, organized, and used. The mass digitization of archives, periodicals and specialized collections, combined with the development of remote access services, has

transformed digital libraries into integrated knowledge management systems that combine information resources, infrastructure and services in a dynamic environment [1, 2]. Recent definitions [3, 4] emphasize that digital libraries integrate technological, organizational, and social dimensions, and are increasingly viewed as environments in which intelligent technologies and artificial intelligence tools provide personalized access and improved user experience [5]. In this sense, a digital library can be understood as a structured information system that stores, indexes, and provides digital resources through network technologies while ensuring their long-term sustainability and accessibility [2].

However, the rapid expansion of digital collections and their growing thematic diversity also introduces significant challenges. When collections contain large numbers of textual resources — particularly periodicals — users are confronted with information overload, as such materials are inherently multi-thematic [6]. A single periodical issue may include political analysis, cultural commentary, sports news, and advertisements that share a common publication context, but lack a unified thematic focus [6]. This internal heterogeneity complicates document modelling and makes it difficult to assess similarity between documents in a way that is both reliable and meaningful for periodicals.

With the growing importance of personalization in digital libraries, numerous studies [3–5] have explored methods and system architectures for delivering personalized access to digital content. Despite this progress, two key challenges remain particularly relevant for periodical collections: identifying effective similarity measures for long, heterogeneous documents and supporting similarity-based access at scale. Given the volume and diversity of digital resources, there is a clear need for approaches in which computationally intensive document representation and similarity estimation are performed offline, while interactive personalization relies on efficient lookup operations over precomputed similarity structures, such as document similarity matrices. In such architectures, similarity assessment must function as a stable and reusable analytical layer that can be computed independently of user interaction and reused across multiple access scenarios. This requirement places specific demands on the design of the similarity model itself, which must be expressive enough to capture meaningful relationships in heterogeneous content while remaining suitable for large-scale, offline computation.

Most existing document similarity models — including those that combine semantic representations with thematic information, metadata, or entity-based features — are designed for relatively short or thematically homogeneous documents and often assume that similarity can be adequately captured through a single global representation. Such assumptions are poorly suited to periodical publications, where similarity frequently arises from partial and localized thematic overlap rather than from overall topical coherence. As a result, relationships between periodicals may depend on limited, but semantically decisive fragments embedded within otherwise dissimilar content, a property that many existing approaches fail to model explicitly.

This work addresses this gap by proposing a multi-component document similarity measure explicitly designed for multi-thematic periodical collections. The novelty of the proposed approach lies not in the introduction of new similarity primitives, but in a similarity formulation that treats partial thematic overlap as a structural characteristic of periodical documents and integrates multiple complementary signals — global semantic context, local fragment-level similarity, thematic distributions, and factual

context — within a single, parameterizable model. By modeling document similarity through the controlled interaction of these components, the proposed measure enables stable and interpretable document-to-document similarity assessment in heterogeneous periodical corpora. From a practical perspective, this formulation supports document-centered, similarity-based navigation in digital library environments, allowing users to locate documents related to the one they are currently accessing.

The remainder of this study is organized as follows: Sect. 2 reviews related work on document similarity, thematic modeling, and similarity assessment in digital library environments. Section 3 presents the proposed multi-component similarity measure and its underlying design principles. Section 4 describes the experimental setup and evaluation methodology and discusses the results and their implications for similarity-based navigation in periodical digital libraries. Section 5 concludes the study and outlines directions for future research.

## 2 Related Work

Digital libraries and information retrieval systems employ several fundamental approaches for identifying similar documents. The earliest and still widely used solutions are based on lexical similarity, where documents are represented by word or term frequencies, and the proximity between them is measured by vector models and measures such as cosine similarity [7, 8]. Such approaches are relatively easy to implement and interpret but suffer from well-known limitations: sensitivity to morphological variations, synonymy and polysemy, as well as an inability to capture deeper semantic relationships between texts [9–11]. Furthermore, if digital resources are multi-thematic periodicals, it is difficult to represent the document through a single theme or short description, and the similarity between two issues often manifests itself locally – at the level of an individual article, column or event – rather than globally across the entire text [5, 12]. The lack of clear relations between documents makes it difficult to navigate the corpus and limits the possibilities for finding similar or thematically related materials [1, 2]. In this context, similarity between documents is increasingly viewed as proximity in semantic space rather than simply as an overlap of lexical units [13–15]. However, even semantic approaches encounter difficulties with periodicals, where much of the content is noisy, repetitive, or strongly tied to specific events and contexts that are not always adequately captured by generalized language models [16, 17]. Along with purely content-based approaches, metadata is often used in practice to assist in the task of finding similar documents [18, 19]. Such metadata may include the date of publication, the edition, the section, the author or predefined thematic classifications. In periodicals, however, metadata is often incomplete, heterogeneous or incompatible between different periods and sources, which limits its reliability and applicability as a standalone source of similarity. Due to these limitations, modern digital libraries are increasingly looking for hybrid solutions that combine different sources of information – content-based, thematic and structured data (e.g. metadata). Particularly in periodicals, it is critically important to capture local thematic matches related to specific events, individuals or institutions that may only be present in limited fragments of the text but are decisive for the semantic proximity between documents.

In this context, the task of finding “similar documents” can be viewed as a multifaceted problem in which similarity is not a one-dimensional quantity, but rather the result of the interaction between global themes, local matches, thematic distribution, and factual context (personalities, places, etc.). Especially in the case of multi-thematic periodicals, the effective modelling of these aspects is a prerequisite for building sustainable, explainable and practically applicable navigation and recommendation services within digital libraries.

### 3 Proposed Multi-component Similarity Measure

#### 3.1 Architectural Overview and Similarity Computation Workflow

The proposed solution is designed to operate in a dynamic digital library environment in which textual resources are continuously added, modified, or removed. To support responsive user interaction over large and heterogeneous corpora, the architecture is organized around a clear separation between computationally intensive similarity estimation and interactive access. In this organization, similarity computation is performed offline, while the interactive layer operates exclusively on precomputed results.

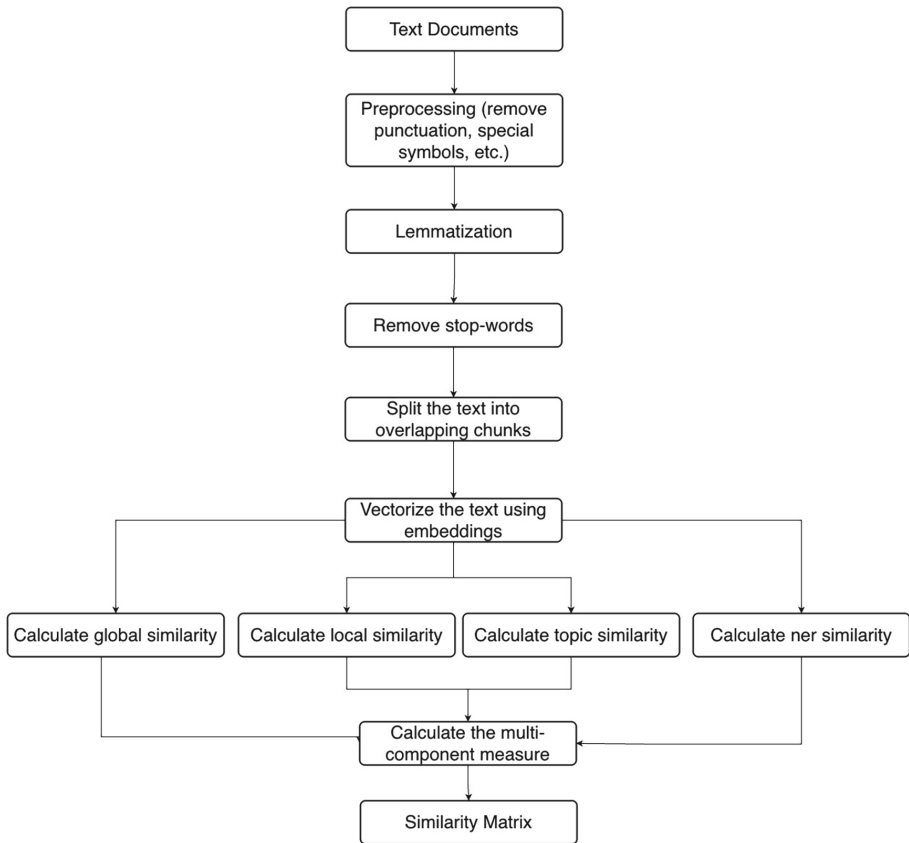
This architectural separation allows similarity assessment to function as a stable and reusable analytical layer, independent of individual user requests. By relying on compact precomputed structures — most notably a document similarity matrix and a shared document index — the interactive layer can efficiently retrieve documents related to the one currently accessed, ensuring short and predictable response times as the collection grows.

During the offline preparation phase, textual resources are processed through a unified workflow that transforms the evolving document collection into compact and reusable similarity structures, as illustrated in Fig. 1 and described in more detail in the following section.

The resulting document similarity matrix constitutes the core operational structure supporting document-centered similarity-based navigation. At runtime, the interactive layer operates exclusively on that precomputed similarity matrix. When a user accesses a document, the system retrieves the corresponding set of related documents by locating the appropriate similarity entries, excluding the active document, and applying simple selection criteria such as thresholds or top-k limits. Because all similarity values are computed in advance, response times remain short and predictable, while the structure of the similarity matrix supports interpretable similarity-based navigation.

To support long-term operation in dynamic collections, document processing and similarity computation are coordinated through a shared system of a persistent document index, which guarantees consistent representation of documents across all internal structures. Changes in the document collection — such as the creation, modification, or removal of resources — trigger updates to the similarity computation workflow. These updates are handled incrementally where possible and may be complemented by occasional full recomputation when necessary to restore global consistency.

The following section presents the formulation of the multi-component similarity measure used to construct the document similarity matrix described above.



**Fig. 1.** The process of generating a similarity matrix using a multi-component measure.

### 3.2 Multi-component Document Similarity Measure

The proposed approach models document similarity in multi-thematic periodical collections using multiple complementary similarity components derived from textual content and named entities extracted from the documents. This section presents the formulation of the multi-component document similarity measure, including document representation, individual similarity components, and their integration into a single similarity score. Named entities are treated as an additional input signal obtained through a separate extraction component of the system architecture; however, the extraction process itself lies outside the scope of the present work.

**Document Representation and Preprocessing.** Reliable measurement of similarity between text documents is directly dependent on the way in which the text is preprocessed, prepared, and subsequently vectorized, as artificial intelligence and machine learning approaches do not work with text, but with numerical values. In the context of digitalized documents that can contain errors from OCR, this requirement is particularly critical, as digitalized resources often contain texts of varying length, style, structure

and quality of digitization. Therefore, the preparation of the data is not a trivial or just technical step, but a conceptual prerequisite for all subsequent document analysis and comparison operations.

The initial text preprocessing step comprises noise removal, text normalization through register standardization, the elimination of special characters and semantically irrelevant elements (e.g., stop words), and lemmatization. Lemmatization reduces lexical items to their base forms and thereby decreases the sparsity of the resulting text representations [20].

Once the text was cleaned to reduce volume and remove semantically uninformative elements (e.g., stop words), the next step is the selection of an appropriate vectorization approach, that is, the transformation of textual data into numerical representations. Classic vectorization approaches, such as “bag of words” and TF-IDF, represent documents as sparse vectors of frequencies or weighted terms and have long been the standard in information retrieval [11]. Despite their transparency and interpretability, these methods model text at a superficial lexical level and do not consider the semantic proximity between different lexical forms. Problems such as synonymy and polysemy lead to situations where documents with similar content are considered distant, and documents with formal overlap in terminology are considered close, regardless of their actual semantic context [11].

Modern natural language processing methods address these limitations by using vector embeddings that encode semantic and contextual information in dense representations [14, 15]. In these models, each word or phrase is represented as a point in a multidimensional space, so that semantically similar units are located close to each other. Unlike static models, which associate the same vector with a given word regardless of its context [21, 22], contextual embeddings position words and sentences according to their specific meaning, allowing for a more nuanced understanding of the meaning and relationships between texts [17]. This property is particularly important when analyzing periodicals, where the meaning of a term often depends on the thematic context. For example, the same lexeme may have a different semantic emphasis in political news, economic analysis, or cultural reviews. Contextual models based on the “transformer” architecture have shown significantly better performance in tasks related to similarity, grouping, and recommendation of texts, precisely because of their ability to encode such dependencies [16, 17].

Therefore, this research uses an approach based on contextual embeddings with a compact representation size, which allows for both high semantic expressiveness and good computational efficiency. The choice of a model is a one published in the publicly available storage Hugging Face [23] with a fixed and relatively low dimension, which eliminates the need for additional dimension reduction techniques such as PCA or LSA [24, 25]. Due to the large volume of data, the texts are processed in batches, and as noted in [26], incremental processing of documents can lead to systemic bias and degradation of representations unless they are periodically retrained. And retraining the model on millions of records cannot be justified, considering the computational overhead involved. Therefore, the chosen model serves two purposes: vectorizing the data by preserving contextual information and reducing the dimensionality of the text.

An additional challenge when working with full-text periodical materials is the limitation on the maximum length of the input sequence imposed by most language models

[16, 17]. Due to the restricted context window of this class of models (typically up to 512 tokens), direct vectorization of long documents using embeddings results in truncation and information loss, which is unacceptable for reliable text analysis.

Although it may seem counterintuitive, providing the maximum-length text as input degrades both processing time and accuracy. To address this issue, each document is segmented into overlapping fragments of controlled length. This widely adopted approach in long-text processing ensures robustness against boundary effects and guarantees that semantic context is preserved across fragment boundaries [16, 17]. Segmentation allows each fragment to be vectorized independently and described through its own semantic representation. In this way, the document is not reduced to a single vector but is modelled as a set of local semantic units that reflect different subtopics and content emphases. The overlap between fragments further mitigates the loss of context at the boundaries and ensures a more stable representation of the entire text [16].

This representation of the document as a set of semantic vectors creates the conditions for defining different complementary similarity measures. On the one hand, by aggregating the fragment vectors, a global similarity can be calculated that reflects the overall thematic orientation of the document. On the other hand, the availability of fragment representations allows for the capture of local similarities between documents, where similarity occurs only in limited parts of the text – a scenario typical for periodicals. Thus, the choice of contextual embeddings in combination with segmentation is not only a matter of representation, but also a structural decision that lays the groundwork for a multi-component similarity assessment.

In this sense, vectorization through contextual embeddings and the segmentation of long documents form a key transition between raw text and the multi-component similarity model. They provide the necessary flexibility for operation and create a stable basis for combining semantic similarity, thematic profiles, and structured signals into a unified assessment of “similar documents”, which will be discussed in the next section.

**Multi-component Measure for Document Similarity.** The task of determining similarity between text documents can rarely be reduced to a single measure. This is particularly true for multi-thematic periodicals, where similarity between two documents may arise at different levels: through a shared overall thematic orientation, through localized overlap of specific articles or sections, through comparable thematic composition, or through references to the same real-world named entities. As noted in several studies on recommendation and retrieval systems [18, 19], relying on a single similarity signal often leads to unstable or difficult-to-interpret results when applied to noisy and heterogeneous corpora.

For this reason, document similarity in this work is modeled using a multi-component measure that integrates several complementary perspectives on relatedness. The measure combines content-based similarity derived from textual representations with an optional factual layer based on named entities extracted from the documents. Each component captures a different aspect of semantic proximity, and together they form a balanced, interpretable, and adaptable similarity assessment.

The content-based similarity between documents  $i$  and  $j$ , denoted as  $S_{\text{content}}(i, j)$ , is defined as a weighted combination of three components: global semantic similarity, local fragment-level similarity, and thematic similarity.

The global component reflects the overall thematic orientation of the documents and is computed by averaging the fragment-level embeddings of each document and measuring cosine similarity between the resulting representations:

$$S_{\text{global}}(i, j) = \cos\left(\dot{c}_i, \dot{c}_j\right) \quad (1)$$

This component provides a stable and robust estimate of similarity at the document level, capturing dominant topics and stylistic coherence [10].

While global similarity offers a useful summary, it inevitably smooths out localized matches that are particularly characteristic of periodicals. Two issues of a newspaper may be related because they share one or more thematically similar articles, even if the remaining content differs substantially. To capture such cases, a local semantic similarity component is introduced, defined as the maximum cosine similarity between all pairs of fragments from the two documents:

$$S_{\text{local}}(i, j) = \max_{k, l} \cos(c_{i, k}, c_{j, l}) \quad (2)$$

This “best match” strategy privileges strong partial overlaps and has proven effective for long and unevenly structured texts, where similarity is concentrated in limited segments rather than distributed uniformly [19].

In addition to semantic similarity at the fragment level, a third component captures similarity in thematic composition through fuzzy clustering. Clustering is an established approach for revealing latent structure in large datasets and is often used as a means of thematically organizing text collections [12, 19, 27]. In its most common forms, it aims to group objects into clearly distinguished sets, with each object belonging to a single cluster [28]. Although such approaches (e.g., k-means) are widely used, they assume a rigid thematic distribution and clearly defined boundaries between groups, which is rarely the case in real text corpora, especially in periodicals with multi-thematic content [28].

To address this limitation, the fuzzy C-means algorithm is employed, allowing fragments to belong to multiple thematic clusters with graded membership coefficients in the interval [0, 1] [29]. Instead of forcing a document into a single category, this approach represents documents as mixtures of themes, reflecting their internal heterogeneity [30]. This property is especially important for periodicals, where a single document may combine political analysis, economic news, cultural commentary, and other genres with varying intensity. Fuzzy clustering allows for precisely this type of interpretation, in which thematic profiles reflect the internal multilayered structure of the content [31].

In the literature, the fuzzy k-means method is considered a suitable tool for working with data characterized by overlapping classes, noise, and uncertainty, such as text corpora in real information systems [28, 31]. Its application in areas such as recommendation systems, user behavior analysis, and social networks shows that “soft” thematic distribution leads to more stable and interpretable models compared to hard classification schemes [29].

In the proposed architecture, fuzzy clustering is not used as a standalone thematic classification tool but as an intermediate analytical layer contributing to similarity



assessment. Fragment-level membership degrees are aggregated into a document-level thematic profile  $t_i$ , and thematic similarity is computed using cosine similarity [30, 31]:

$$S_{\text{topic}}(i, j) = \cos(t_i, t_j) \quad (3)$$

This component captures the similarity in the distribution of thematic weights and allows documents that share a combination of themes to be marked as similar, even when none of them is dominant [29, 30]. This thematic profile is used as a complementary component in calculating the similarity between documents, along with semantic proximity and factual signals. In this way, thematic information does not dominate the assessment but acts as a stabilizing and explanatory factor that helps to distinguish between documents with a similar overall style but different thematic distribution, as well as between documents that share a combination of topics without matching entirely in content [28, 29].

The use of fuzzy clustering as part of a multi-component similarity measure is particularly suitable for multi-thematic periodicals, where the goal is not strict classification but rather the capture of partial, and contextually conditioned similarities. In this sense, the method provides a conceptual framework that allows the thematic structure of the corpus to be integrated into a broader model for content analysis and recommendation [29–31].

The influence of the individual similarity components is controlled through an explicit parameterization, allowing the contribution of each aspect of similarity to be adjusted independently. The three content-based components are combined linearly:

$$S_{\text{content}}(i, j) = \alpha S_{\text{global}}(i, j) + \beta S_{\text{local}}(i, j) + \gamma S_{\text{topic}}(i, j), \alpha + \beta + \gamma = 1 \quad (4)$$

These parameters regulate the relative influence of global context, localized overlap, and thematic composition. Increasing  $\alpha$  emphasizes dominant themes and stylistic coherence, higher values of  $\beta$  prioritize strong partial matches between specific fragments, and  $\gamma$  controls the contribution of thematic structure. When thematic modeling is not applied,  $\gamma$  is set to zero.

Semantic similarity models may overlook relationships driven by rare or specific named entities such as persons, institutions, or locations, which often carry decisive informational value in periodical texts. To address this limitation, an additional factual similarity component based on named entities is introduced. For each document  $i$ , a set  $E_i$  of extracted named entities is defined, and similarity is measured using the Jaccard coefficient:

$$S_{\text{NER}}(i, j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|} \quad (5)$$

The final similarity score is obtained as:

$$S_{\text{final}}(i, j) = (1 - \lambda) S_{\text{content}}(i, j) + \lambda S_{\text{NER}}(i, j), 0 \leq \lambda \leq 1 \quad (6)$$

The parameter  $\lambda$  controls the balance between semantic and factual similarity. When  $\lambda = 0$ , similarity is determined exclusively by content-based signals, while increasing  $\lambda$  strengthens the influence of shared named entities. In periodical collections, this parameter acts as a corrective and explanatory factor, reinforcing similarity when documents refer to the same real-world objects even if their semantic framing differs.

This formulation creates conditions for experimental calibration of the parameters depending on the specifics of the corpus [18, 29, 30]. Overall, the proposed multi-component formulation allows similarity to be modeled as the interaction of multiple interpretable signals rather than relying on a single dominant signal [28–30]. This structure enables explicit control over different aspects of document relatedness and is particularly well suited to multi-thematic periodicals, where similarity is often partial, contextual, and factually grounded.

## 4 Experimental Setup, Results and Discussion

The empirical evaluation of the proposed multi-component similarity measure is non-trivial, as real-world systems typically lack a predefined gold standard specifying which documents should be considered similar. Consequently, there is no direct reference against which the generated similarity matrix, produced by the measure, can be systematically validated. This is particularly true for corpora of periodicals, where thematic boundaries are blurred, and similarities are often partial and context dependent. Such limitations are widely discussed in the literature on recommendation systems and information retrieval, where the need to combine quantitative and qualitative evaluation methods is emphasized [18, 19].

In view of these peculiarities, experimental validation is organized as a multi-step process that combines parametric optimization, ablation analysis, and qualitative interpretation of the results. The main goal is not to achieve a maximum value for a specific metric, but to assess the robustness, interpretability, and practical usefulness of the similarities generated by the model in the context of a multi-thematic and noisy corpus.

As mentioned, there are no explicit similarity labels, so a reference matrix (“gold standard”) is constructed that combines two independent signals: 1) overlap of lexical units after lemmatization and synonym enrichment, and 2) overlap of named entities extracted from the documents. Such a multi-component reference approach is in line with practices for internal validation of similarity models, which use weakly supervised or proxy metrics instead of manual annotations [9, 10].

To assess the robustness of the results with respect to corpus size and diversity, all experiments are conducted on samples of increasing scale: small corpora ( $5 \times 200$  documents), medium-sized corpora ( $2 \times 500$  documents), and a control corpus comprising a larger volume of documents (1,000). This experimental design enables an analysis of how model behavior evolves as the amount of data increases, and whether the optimal parameter configurations remain stable across scales — an essential consideration for systems intended to support incremental corpus expansion [17].

Although formal statistical significance testing is difficult in the absence of explicit relevance judgments, this experimental design provides an indirect assessment of result stability. By repeating the evaluation across independent samples and progressively larger corpora, it becomes possible to observe whether performance differences persist consistently or fluctuate due to random variation. The consistency of observed trends across scales is therefore treated as an indicator of robustness rather than relying on single-run measurements.

The first step of the experimental procedure for validating the effectiveness of the multi-component measure involves a systematic investigation of the parameter space  $\alpha$ ,  $\beta$ , and  $\gamma$ , which determine the relative weight of the individual components in the final score of similarity. Such an approach, known as grid search, is a standard method for calibrating parameterized models when there is no analytical solution for the optimal values [18]. For each parameter configuration, a similarity matrix is computed and subsequently evaluated with respect to the extent to which it reproduces the expected inter-document relationships, as defined by the constructed gold standard.

The results of the parametric optimization show a clearly distinguishable role for the individual components. The parameter  $\alpha$ , which controls the balance between content similarity and similarity by named entities, reaches an optimal value of around  $\alpha = 0.5$ . This means that the semantic and factual signals contribute with approximately equal weight to the final score. The result can be explained by the specifics of periodicals, in which the informational value is often concentrated around specific individuals, institutions, and events that are not always captured reliably enough by purely semantic models [10].

The analysis of the parameters  $\beta$ ,  $\gamma$ , and  $\delta$ , which determine the internal structure of the content component, reveals additional systematic patterns. Notably, the optimal parameter values are invariant with respect to sample size, with identical configurations observed across all experimental settings. The highest weights are consistently assigned to global semantic similarity ( $\beta = 0.47$ ) and thematic similarity through fuzzy clustering ( $\gamma = 0.47$ ), whereas the contribution of local similarity ( $\delta$ ) remains relatively low. This shows that the overall thematic context and distribution of topics are more reliable indicators of similarity between documents than individual fragment matches, which can often be the result of noise, standard headings, advertisements, or repetitive wording in periodicals [30, 31].

As a second step, the marginal contribution of each component is assessed through an ablation analysis (see Table 1), in which the individual components of the multi-component measure are sequentially removed. Such analyses are widely used for diagnosing complex models and enable the identification of essential architectural components as opposed to secondary ones [29].

**Table 1.** Ablation study results

Sample size	Global similarity	Content similarity	Final similarity
200	0.1022	0.0712	0.0959
200	0.1109	0.0927	0.1061
200	0.0884	0.0831	0.0918
200	0.1021	0.08	0.0942
200	0.0937	0.0665	0.0927
500	0.1067	0.0808	0.0954
500	0.0984	0.0772	0.0911
1000	0.1025	0.0776	0.0895

Across all corpus sizes and repeated samples, the relative ordering of configurations remains stable: the full multi-component configuration consistently outperforms both the global-only and content-only variants. While absolute metric values vary slightly between samples, the direction and magnitude of improvements are preserved, indicating that the observed differences reflect systematic effects of the model components rather than incidental fluctuations in the data.

The results reveal a consistent and characteristic pattern in the behavior of the quality metrics. Specifically, increasing model complexity initially leads to a decline in the lexical overlap metric, followed by a marked recovery upon full integration of the different components. This trend indicates that the base model primarily maximizes coverage, whereas the complete multi-component configuration enhances precision and robustness. Consequently, the selected configuration, which assigns equal weight to the semantic and factual layers, is particularly well suited to periodic multi-topic documents, where the reliability and interpretability of inter-document relations are of primary importance. This recovery can be attributed to the incorporation of similarity signals derived from named entities, which reintroduce connections only when documents share a concrete factual context. Unlike purely semantic generalization, this component selectively restores links that are grounded in explicit references, thereby explaining the observed improvement in precision without a corresponding loss of interpretability [10, 18].

Additionally, a controlled experimental verification was conducted on a synthetic corpus constructed to contain clearly defined thematic groups, partial overlaps, and shared named entities. This type of experiment allows for a deterministic definition of expected similarities and is an established approach for testing the sensitivity of models to different types of signals [9]. The results confirm that the full multi-component measure is more successful in detecting partial and contextual connections compared to the basic configurations.

In the context of similarity-based navigation over large periodical collections, similarity scores directly determine which documents appear among the top-k neighbors of a given document. Even small improvements can therefore result in different documents being included or excluded from recommendation sets, particularly near selection thresholds. In practice, this translates into fewer spurious links between unrelated documents and more coherent groupings of thematically related materials, which aligns with the qualitative behavior observed in the ablation analysis.

In summary, experimental validation shows that the proposed multi-component similarity measure is robust to data scale, sensitive to different aspects of similarity, and amenable to interpretable calibration through a limited set of parameters. The analysis of the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  demonstrates that the balanced combination of semantic, thematic, and factual signals is particularly suitable for multi-thematic periodicals and provides a reliable basis for building practical services for finding “similar documents” for periodicals [29–31]. Importantly, these improvements remain consistent across corpus sizes and parameter configurations, reinforcing the practical reliability of the proposed approach for similarity-based navigation in periodical digital libraries.

## 5 Conclusion and Future Work

This study demonstrates that personalized access to digital library collections dominated by periodical publications requires document similarity models that explicitly accommodate thematic heterogeneity, partial overlap, and context-dependent relatedness. The proposed multi-component document similarity measure is well-suited to multi-thematic periodicals, as it models similarity as the combined effect of multiple complementary signals rather than as a single, one-dimensional score. By integrating semantic, thematic, and factual aspects of relatedness, the approach supports more stable and interpretable identification of documents related to a given item within heterogeneous collections.

Beyond its immediate application, the work has broader implications for digital libraries and information retrieval research. Conceptually, it highlights the limitations of global, single-representation similarity models when applied to long and internally diverse documents and illustrates the value of treating document similarity as a composite phenomenon. From a digital library perspective, the results suggest that similarity-based navigation and personalization services can be made more robust by grounding them in similarity measures explicitly designed for the structural properties of periodical content. Practically, the formulation supports deployment in large-scale environments by enabling similarity computation to be performed offline and reused across multiple access and navigation scenarios.

Future work will focus on extending the proposed measure with additional contextual signals and exploring its integration with user interaction data to support adaptive personalization strategies. More broadly, the approach provides a foundation for developing sustainable similarity-based access services that remain effective as digital library collections grow in size, diversity, and thematic complexity.

**Acknowledgment.** This research builds upon a software library developed with the partial funding and support by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, <https://clada-bg.eu/bg/>, Grant number DO01-97/26.06.2025, financed by the Bulgarian Ministry of Education and Science, Funding Procedure: The National Science Infrastructure Roadmap 2020–2027 [Bul20] BULGARIAN MINISTRY OF EDUCATION AND SCIENCE. “National Science Infrastructure Roadmap 2020–2027”. [https://web.mon.bg/upload/26649/RoadMapBulgaria\\_2020-2027\\_BG\\_sm\\_11062021.pdf](https://web.mon.bg/upload/26649/RoadMapBulgaria_2020-2027_BG_sm_11062021.pdf).”

## References

1. Liu, Z., Shao, B.: A systematic review of library services platforms research and research agenda. *Libr. Inf. Sci. Res.* **46**(4), 101325 (2024). <https://doi.org/10.1016/j.lisr.2024.101325>
2. Borgman, C.L.: Libraries, digital libraries, and data: forty years, four challenges. *Portal: Libraries and the Academy.* **25**(3), 39–58 (2025). <https://doi.org/10.1353/pla.2025.a964199>
3. Fekadu, M., Alemneh, D.: Digital library models: a systematic review. In: *International Conference on Asian Digital Libraries*, pp. 87–101. Springer Nature Singapore (2025). [https://doi.org/10.1007/978-981-96-0865-2\\_7](https://doi.org/10.1007/978-981-96-0865-2_7)

4. Owusu-Ansah, C.M., Rodrigues, A.D.: Digital information and library services in ODDE: towards a collaborative digital library model. In: *Handbook of Open, Distance and Digital Education*, pp. 819–839. Springer Nature Singapore (2023). [https://doi.org/10.1007/978-981-19-0351-9\\_45-1](https://doi.org/10.1007/978-981-19-0351-9_45-1)
5. Kumar, N.: The role of artificial intelligence in enhancing digital library services. *Int. Res. J. Libr. Inform. Sci.* **2**(8), 9–17 (2025)
6. Neupane, A., Khanal, K., Nepal, N., Dangi, N.: Advanced news aggregation and content generation using LLMs and NLP algorithms. *European J. Appl. Sci. Eng. Technol.* **3**(2), 295–303 (2025). [https://doi.org/10.59324/ejaset.2025.3\(2\).24](https://doi.org/10.59324/ejaset.2025.3(2).24)
7. Prasetya, D.D., Wibawa, A.P., Hirashima, T.: The performance of text similarity algorithms. *Int. J. Adv. Intell. Inf.* **4**(1), 63–69 (2018). <https://doi.org/10.26555/ijain.v4i1.152>
8. Zhang, W.: A practical algorithm for efficiently deduplicating highly similar news in large news corpora. In: *CS & IT Conference Proceedings*, vol. 13 (2023)
9. Jurafsky, D., Martin, J. H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 1st edn. (2020)
10. Levy, A., Shalom, B.R., Chalamish, M.: A guide to similarity measures and their data science applications. *J. Big Data.* **12**(1), 188 (2025). <https://doi.org/10.1186/s40537-025-01227-1>
11. Ha, T., Gao, X.: Evolving multi-view autoencoders for text classification. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 270–276 (2021). <https://doi.org/10.1145/3486622.3493969>
12. Lu, Y., Xin, H., Wang, R., Nie, F., Li, X.: Scalable multiple kernel k-means clustering. In: *31st ACM International Conference on Information & Knowledge Management*, pp. 4279–4283 (2022). <https://doi.org/10.1145/3511808.3557690>
13. Mersha, M.A., Kalita, J.: Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. *Procedia Computer Science.* **121-132**, 10.48550/arXiv.2410.00134 (2024)
14. Patil, R., Boit, S., Gudivada, V., Nandigam, J.: A survey of text representation and embedding techniques in NLP. *IEEE Access.* **11**, 36120–36146 (2023). <https://doi.org/10.1109/ACCESS.2023.3266377>
15. Nie, Z., Feng, Z., Li, M., Zhang, C., Zhang, R.: When text embedding meets large language model: a comprehensive survey. *arXiv preprint arXiv, 2412.09165* (2024). <https://doi.org/10.48550/arXiv.2412.09165>
16. Alkaabi, H., Jasim, A.K., Darroudi, A.: From static to contextual: a survey of embedding advances in NLP. *PERFECT: J. Smart Algorithms.* **2**(2), 64–73 (2025). <https://doi.org/10.62671/perfect.v2i2.77>
17. Das, K., Abid, F.: Advancements in word embeddings: a comprehensive survey and analysis. In: *Proceedings of the Pakistan Academy of Sciences: a. Physical and Computational Sciences*, vol. 61, pp. 227–245 (2024). [https://doi.org/10.53560/PPASA\(61-3\)842](https://doi.org/10.53560/PPASA(61-3)842)
18. Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A.: Recommendation systems: algorithms, challenges, metrics, and business opportunities. *Appl. Sci.* **10**(21) (2020). <https://doi.org/10.3390/app10217748>
19. Liqiang, H., Quan, L.: Design of resource recommendation model for personalized learning in the era of big data. In: *AMME 2019: Proceedings of the 2019 Annual Meeting on Management Engineering*, pp. 181–187 (2019). <https://doi.org/10.1145/3377672.3378054>
20. Wibawa, A.P., Kurniawan, F.: Advancements in natural language processing: implications, challenges, and future directions. *Telematics Inform. Rep.* **16**, 100173 (2024). <https://doi.org/10.1016/j.teler.2024.100173>
21. Zhang, C., et al.: From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models. *arXiv:2411.05036*. (2024). <https://doi.org/10.48550/arXiv.2411.05036>

22. Golec, J., Hachaj, T.: Ten natural language processing tasks with generative artificial intelligence. *Appl. Sci.* **15**(6), 9057 (2025). <https://doi.org/10.3390/app15169057>
23. Hugging Face.: <https://huggingface.co> Last accessed 09 Jan 2026
24. Ashraf, M., et al.: A survey on dimensionality reduction techniques for time-series data. *IEEE Access.* **11**, 42909–42923 (2023). <https://doi.org/10.1109/ACCESS.2023.3269693>
25. Chang, Y.C.: A survey: potential dimensionality reduction methods. *arXiv preprint arXiv*, 2502.11036 (2025). <https://doi.org/10.48550/arXiv.2502.11036>
26. Shahzad, M., Barzamini, H., Wilson, J., Alhoori, H., Rahimi, M.: Dynamic domain analysis for predicting concept drift in engineering AI-enabled software. *J. Data Inform. Sci.* **10**(2), 124–151 (2025)
27. Mehta, V., Bawa, S., Singh, J.: WEClustering: word embeddings based text clustering technique for large datasets. *Complex Intell. Syst.* **7**(6), 3211–3224 (2021). <https://doi.org/10.1007/s40747-021-00512-9>
28. Lu, H., Bao, Y., Gao, Q.: Distance-based fuzzy K-means clustering without cluster centroids. *Signal Process.* **241**, 110406 (2026). <https://doi.org/10.1016/j.sigpro.2025.110406>
29. Li, Y., Xie, X.: Deep multi-view fuzzy k-means with weight allocation and entropy regularization. *Appl. Intell.* **53**(24), 30593–30606 (2023). <https://doi.org/10.1007/s10489-023-05113-2>
30. Ferraro, M.B.: Fuzzy k-means: history and applications. *Econom. Stat.* **30**, 110–123 (2024). <https://doi.org/10.1016/j.ecosta.2021.11.008>
31. Geng, X., Mu, Y., Mao, S., Ye, J., Zhu, L.: An improved K-means algorithm based on fuzzy metrics. *IEEE Access.* **8**, 217416–217424 (2020). <https://doi.org/10.1109/ACCESS.2020.3040745>