



**BULGARIAN ACADEMY OF SCIENCES
INSTITUTE OF MATHEMATICS AND INFORMATICS**

Emanuela Dimitrova Mitreva

**METHODS AND ALGORITHMS FOR
PERSONALISATION AND ADAPTABILITY IN
CONTENT MANAGEMENT SYSTEMS**

Abstract

of a dissertation

for awarding educational and scientific degree PhD

in the field of higher education 4. Natural Sciences, Mathematics and Informatics,
professional field 4.6. Informatics and Computer Sciences, doctoral program “Informatics”

Scientific supervisor:

Professor Desislava Paneva-Marinova, PhD

Sofia, 2026

The dissertation was discussed and referred for defence at an extended meeting of the Mathematical Linguistics Section at the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences on 14.09.2026. The dissertation is **167** pages long and contains **15** tables and **26** figures. It includes an introduction, **5** chapters, a list of **202** references, and a list of 5 publications by the author (1 of which is independent) related to the dissertation.

The numbering of the tables and figures in the abstract follows the original numbering used in the dissertation.

The materials for the defence are available to interested parties at the Institute of Mathematics and Informatics - BAS, ul. "Akad. G. Bonchev", block 8, Sofia.

Author: Emanuela Dimitrova Mitreva

Title: Methods and Algorithms for Personalisation and Adaptability in
Content Management Systems

CONTENTS

Chapter 1. General characteristics of the dissertation	4
1.1. Relevance of the problem.....	4
1.2. Object, scope, aim, and objectives of the study	4
1.3. Structure of the dissertation.....	6
Chapter 2. Theoretical foundations and analysis of contemporary approaches to personalization in digital libraries.....	8
2.1. Digital libraries: concept and evolution	8
2.2. Personalisation in digital libraries	9
Chapter 3. Models and software components for personalized content presentation in digital libraries	15
3.1. Conceptual model and architectural framework for personalized content presentation in a digital library	15
3.2. Service for extracting and structuring named entities	16
3.3. Similarity matrix and the method of multi-component similarity assessment	17
3.4. User-document matrix and implicit assessment	19
3.5. Operational structures and mechanisms for updates	20
3.6. Modules for recommending personalised content.....	21
3.7. Explainability and ethical principles in the selection of personalised content	22
Chapter 4. Experimental implementation and analysis of performance testing	24
4.1. Building a technological environment, test data and protocol for experimental verification	24
4.2. System architecture	25
4.3. Service for extracting and structuring named entities	26
4.4. Similarity matrix and multi-component assessment method. Functional module for selecting “similar documents”	27
4.5. Sparse “user-document” matrix, hybrid algorithm, and functional module for generating “personalized recommendations”	29
4.6. Limitations and validity of the proposed architecture	30
CONTRIBUTIONS OF THE DISSERTATION	31
APPROBATION	33
List of publications on the topic of the dissertation	33
List of reported results	34
List of citations.....	35
Bibliography.....	36

CHAPTER 1. GENERAL CHARACTERISTICS OF THE DISSERTATION

1.1. Relevance of the problem

In recent decades, digital transformation has significantly changed the way scientific and cultural heritage is stored and used. The accumulation of large amounts of digitized resources and the remote access to them have made digital collections an integral part of scientific activity and education. In this context, digital libraries occupy a special place as environments that combine long-term storage, reliable description, and organised provision of information resources that are diverse in origin and structure.

However, with the growth in volume and diversity of content, a significant challenge arises: standard search and navigation approaches are finding it increasingly difficult to help users find documents that actually match their needs or interests. The abundance of resources, the lack of explicit ratings, and varying degrees of structure often lead to too many results and difficulties in navigating the available resources. This highlights the need for approaches that not only provide access to resources but also make it more selective and tailored to the needs of the specific user.

In this sense, personalised content is seen as a promising direction for the development of digital libraries. Using text analysis, structured descriptions, and logs of interactions between users and documents makes it possible for modern solutions in the field of artificial intelligence to support the discovery of semantically similar resources while complying with requirements for transparency and the protection of personal data. Of particular interest are hybrid solutions that combine multiple sources of information and different approaches to achieve more robust, explainable, and practically applicable recommendations in real-world environments.

1.2. Object, scope, aim, and objectives of the study

The object of the dissertation is the process of adapting and personalising content in digital libraries using artificial intelligence and machine learning methods and techniques. The research focuses on ways in which digital libraries can analyse user behaviour, preferences, and context to provide dynamic content tailored to the individual needs and interests of each user.

The dissertation examines and systematises contemporary approaches to provide personalised content in digital libraries based on artificial intelligence and machine learning methods, outlining the main problems and challenges in their implementation. The **main aim** is to develop new models, methods, and tools for providing personalised content that combine

content characteristics, data from user-document interaction logs, and metadata. The aim is to offer users the most relevant, understandable, and tailored information resources while maintaining scalability and proven applicability in a real environment.

The scope of the study is approaches, models, and algorithms for adapting information objects and resources in digital libraries in order to provide personalised content.

The research is based on the hypothesis that the application of appropriate methods of artificial intelligence and machine learning for adapting content in digital libraries leads to a higher degree of personalisation, relevance, and efficiency in providing information to users.

In line with this hypothesis and with a view to achieving the aim of the dissertation, the following main research objectives have been formulated:

Objective 1: To study the scientific achievements and results of current research on the use of artificial intelligence and machine learning methods for providing personalised content in digital libraries.

Objective 2. To explore the possibilities for applying modern methods of artificial intelligence and natural language processing using large language models to extract named entities from text resources and integrate them as structured metadata in order to improve search capabilities and use them as an additional information source in the construction of hybrid recommendation modules.

Objective 3. To create a conceptual model of functional modules for recommending content in a digital library, based on modern methods of artificial intelligence, which offers both information resources semantically similar to those currently being viewed and other resources that the user might potentially be interested in. Within the model, develop approaches for processing and using user data, as well as approaches for increasing the transparency and explainability of the recommendation process.

Objective 4. Develop and implement a prototype of the proposed functional modules and conduct experimental testing to assess their effectiveness and applicability.

Objective 5. Analyse and interpret the results of the experiments conducted in order to draw conclusions about the quality of the recommendations and the potential of the proposed module.

1.3. Structure of the dissertation

The structure of the dissertation is as follows:

Chapter 1. General Problem Statement defines the research object and scope, the primary aim and specific objectives, and outlines the context in which the personalized presentation of content in digital libraries is examined.

Chapter 2. Theoretical Foundations and Analysis of Contemporary Approaches to Personalization in Digital Libraries presents the key concepts, models, and classifications related to personalization and content recommendation algorithms, and conducts an analytical review of scientific achievements and results from recent research on the application of artificial intelligence and machine learning methods for delivering personalized content in digital libraries.

Chapter 3. Models and Software Components for Personalized Content Presentation in Digital Libraries formulates the theoretical framework of the proposed architecture for personalized content presentation in a digital library. The chapter introduces the system's conceptual model and the roles of its core components-two functional modules for generating personalized content and a separate service for the extraction and structuring of named entities. It systematically describes the stages of data preparation and the creation of operational structures within the asynchronous layer. Based on these structures, two principal approaches to personalized content delivery are defined in the interactive layer: (1) the discovery of similar documents, invariant across users; and (2) personalized recommendations, in which personalization is achieved through a hybrid approach that combines content similarity to previously accessed resources with a global popularity indicator, thereby balancing individual user preferences and stable trends in the absence of behavioural history. The exposition reveals the interdependencies among the modules and justifies the choice of a hybrid approach.

Chapter 4. Experimental Implementation and Analysis of Testing Results presents the practical implementation of the proposed modules and components, the software tools and parameters employed, the evaluation methodology, and the results of experimental testing, through which the applicability and effectiveness of the proposed solutions are assessed in a real-world digital library environment.

Chapter 5. Conclusion and Future Directions summarizes the results achieved through the development, analysis, and experimental implementation of the proposed solutions,

confirms their effectiveness and applicability in the context of digital libraries, and outlines potential directions for future work, including the extension of functionality, performance optimization, and integration with additional standards and intelligent methods for the management and analysis of digital content.

CHAPTER 2. THEORETICAL FOUNDATIONS AND ANALYSIS OF CONTEMPORARY APPROACHES TO PERSONALIZATION IN DIGITAL LIBRARIES

The rapid development of information technology in recent decades has led to fundamental changes in the way knowledge is created, stored, and shared. The exponential growth of digital resources and the increasing accessibility of information have transformed traditional approaches to knowledge management and created a need for new methods of organising and using it. However, this transformation poses new challenges for researchers and developers in terms of the effective management of information resources and ensuring meaningful access to knowledge in conditions of information overload.

This chapter aims to present theoretical foundations and contemporary concepts for the personalisation of content, focusing on the principles, methods, and algorithmic approaches that define the development of this research area. Established and innovative solutions based on artificial intelligence and machine learning techniques are examined, enabling the adaptation of the information environment to individual user interests and behavioural patterns. Within the scope of the analysis, the evolution of personalization approaches and their applicability across different contexts is traced, with particular emphasis placed on the examination of their implementations in the domain of digital libraries.

2.1. Digital libraries: concept and evolution

Digital libraries are perceived not merely as digitized collections, but as integrated knowledge management systems that combine information resources, infrastructure, and services within a dynamic environment [1], [2]. In this sense, the main goal of digital libraries is to provide broad, sustainable and equal access to knowledge by supporting scientific and cultural processes on a global scale [2], [3]. They aim not only at the digitization of content, but also at the development of an intelligent information ecosystem that brings together data, services, and users within a dynamic environment of interaction [2], [4], [5].

The authors of [6], [7] add that through the adoption of artificial intelligence and adaptive technologies, digital libraries achieve a higher degree of personalization and accessibility, thereby facilitating navigation and enhancing the efficiency of research activities. In this way, their importance goes beyond the traditional function of storing and providing information – they become an active mediator in the creation, discovery, and sharing of knowledge, supporting the development of innovation, academic practices, and cultural

memory, and serve as an active environment for the creation, sharing, and development of knowledge, which is at the heart of contemporary scientific and educational processes [2], [6].

The digital environment also changes the way users interact with information resources by creating conditions for personal access and dynamic search. According to [8], digital libraries are no longer just repositories of information, but "adaptive platforms" that analyse user behaviour patterns in order to provide more relevant content and an improved user experience. Personalisation is a key approach to improving the effectiveness of information interaction. It involves adapting content, interface and functionalities to the individual characteristics, interests and behaviour of the user [9]. The user is no longer just a static consumer of information and resources, but an active participant in the construction of knowledge.

As a result, the need for personalised solutions in digital libraries is seen as a necessary change to improve user satisfaction and the effective use of information resources. Digital libraries offering personalised content not only help users find relevant content but also create a more intelligent and engaging learning and research environment that reflects digital knowledge – user-oriented, context-oriented, and interaction-oriented.

2.2. Personalisation in digital libraries

The era of big data has established information as a strategic resource, and the ability to extract knowledge from large and heterogeneous data sets has become a key competitive advantage [10]. However, the growing volume and complexity of data make it difficult to analyse and navigate information [2], which requires the development of intelligent decision support systems tailored to the individual needs of the user [2], [11]. Although modern methods of data analysis and machine learning allow the discovery of hidden patterns, their effectiveness depends on the context and quality of the data used [12].

In digital libraries, personalisation is particularly important due to the large volume of digitised documents and resources. Intelligent recommendation systems analyse content and user behaviour, providing more accurate classification and access to relevant information and increasing user engagement [13], [14].

Personalization approaches may be static, based on predefined settings, keywords, or previously conducted user surveys [15]. Simple statistical methods can also be used, such as generating a list of items that are of interest to most users or that reflect a particular area of interest. The other type of personalisation approaches is dynamic and uses recommendation

algorithms that analyse large data sets and adapt content in real time. Their successful application on leading platforms such as YouTube, Amazon, and Netflix demonstrates their potential, and their transfer to digital libraries is a logical step towards more efficient and interactive organisation and provision of knowledge.

2.2.1. Modern methods, algorithms, and personalization approaches

Contemporary approaches that provide personalised content are based on adaptive methods, which in most cases include recommendation systems, user behaviour analysis, or a combination of both approaches. They are also based on the idea that the information system must continuously adapt to the changing needs, interests, and behaviour of the individual user. These methods use an integrated set of artificial intelligence and machine learning techniques and algorithms that work together to provide dynamic and relevant content delivery.

The main approaches to generating personalized content are based on user profiling and behavioural analysis, as well as on classical recommender system methods-content-based filtering, collaborative filtering, and hybrid solutions. These approaches are often complemented by classification and clustering methods, which support the identification of similarities and patterns and improve the accuracy of recommendations.

The combination of behavioural analysis, recommendation algorithms, and analytical methods forms adaptive and learning systems capable of both responding to current user needs and predicting future interests. Despite existing limitations related to data quality and completeness, these approaches provide an effective basis for building personalised information environments with increased utility and engagement.

The first approach considered is **web usage mining**, which uses data from user interactions with the system, such as search history, navigation, and content ratings. By processing and modelling this data, user profiles are built that serve as input for recommendation algorithms and enable the tailored presentation of relevant content, based on individual interests and context.

In addition to approaches based on profiling and user behaviour analysis, classic recommendation systems are widely used: content-based filtering and collaborative filtering.

Content-based filtering uses the similarity between objects and the individual user profile, and its effectiveness depends on the availability of sufficient information about the user, as well as on the quality of the metadata and semantic links between resources [16]. A major

limitation of this approach is its dependence on the user's already known interests and the difficulty of discovering new content outside the established context. The accuracy of recommendations can be improved by including user ratings or indirect indicators of interest, such as time spent interacting with the content, but the method remains limited in systems with rich and poorly structured information spaces [17], [18].

Collaborative filtering overcomes some of these limitations by building recommendations based on similarities between users or between elements extracted from user behaviour patterns. Analysing collective preferences provides more effective and scalable personalisation, especially when sufficient data is available. Despite the challenges associated with processing large volumes of information and selecting appropriate similarity metrics, this approach is establishing itself as a key mechanism for delivering personalised content in adaptive information systems [17].

In the context of personalised systems and adaptive recommendation approaches, clustering and classification methods play a key role, even if only as auxiliary methods. There are three main training methods – **supervised training**, **unsupervised training** and **semi-supervised training**, which differ both in the availability and volume of preliminary information about the data and in the way the dependencies between the input and output variables are modelled.

Supervised learning is based on pre-classified data and is widely applied to classification and regression tasks, including natural language processing and thematic text categorization. This approach enables the development of models with high predictive accuracy; however, its primary limitation is the requirement for large and reliably labelled training datasets, the creation of which is often resource-intensive [19]. Among the most commonly used algorithms are the Naive Bayes classifier, support vector machines, and the k-nearest neighbours (KNN) method, which are characterised by high accuracy in predictive tasks [20], [21], [22]:

- **The Naive Bayesian Classifier** is a probabilistic classification method based on Bayes theorem and the assumption of conditional independence between features [23]. Despite the simplifying assumption, the algorithm shows good efficiency and high computational speed, especially in text classification and natural language processing tasks [23]. The advantages of the approach include easy implementation, scalability, and stable performance on large and multidimensional datasets.

- **Support Vector Machine (SVM)** is one of the most popular approaches to supervised machine learning, in which no prior knowledge of the problem domain is used [24]. SVM works very well with high-dimensional data, avoiding the "curse of dimensionality" [24]. Only a portion of the training examples are used – so-called support vectors – to represent a decision surface [24]. The problems that SVM solves are usually classic classification tasks, where we have two classes [24].
- The **k-nearest neighbours (k-NN)** method is a non-parametric and "lazy" approach in which no explicit model is constructed: for a new object, the k nearest examples in the training set are found according to a selected metric, and the solution is obtained by voting (in classification) or averaging/weighted averaging (in regression) [25], [26]. The effectiveness of k-nearest neighbours depends directly on the representation of the data, the choice of proximity measure and the value of k [27].

Clustering is an **unsupervised learning** technique which goal is to group data into similar clusters without predefined labels, thereby discovering internal structures and patterns in the data set [28]. One of the most widely used methods is k-means clustering, which partitions objects into a predefined number of clusters by iteratively minimizing intra-cluster variance. Its effectiveness is highly dependent on the choice of the number of clusters. For data with ambiguous or overlapping group boundaries, a natural extension of this approach is the fuzzy k-means method, which allows graded membership of objects in more than one cluster and provides more flexible and realistic modelling of complex, noisy, and dynamic data characteristic of adaptive and personalized systems [29].

An intermediate and increasingly widely adopted approach is semi-supervised learning, which combines a limited set of labelled data with a large volume of unlabelled information [30]. This strategy reduces dependence on costly annotation processes and provides a better balance between accuracy and applicability, particularly in domains such as document classification and natural language processing.

2.2.2. Text representations and similarity measures for personalization

In the context of personalization, particularly when applying machine learning and artificial intelligence methods, not only the selected algorithm but also the choice of an appropriate text representation and similarity measure is of critical importance, as data processing and analysis require information to be expressed in numerical form, which

necessitates the application of vectorization techniques and the formalization of textual information.

The established and widely adopted approaches to vectorisation are:

- **CountVectorizer** converts text data into numerical form by building a dictionary of unique terms and calculating the frequency of occurrence of each term in each document [31]. The result is a sparse "document-term" matrix that can be extended with n-gram features to capture local dependencies [31]. The method is fast, transparent, and suitable as a base model, but it does not take into account semantic relationships between terms, so it is often combined with dimension reduction techniques.
- **HashingVectorizer** is a feature extraction technique in which tokens are mapped via a hashing function to indices in a pre-fixed space, resulting in a sparse frequency matrix [32]. Unlike count-based representation, the method does not maintain an explicit dictionary, which provides better scalability and memory efficiency and makes it suitable for large or streaming data [32]. The main limitations of the approach are the inability to reverse interpret the features and the risk of hash collisions, which can be reduced but not completely eliminated by choosing a sufficiently large hash space [32].
- **TF-IDF, TfidfVectorizer** is a method for numerically representing text that evaluates the significance of a term in relation to a specific document and the entire collection by combining the frequency of the term in the document (TF) and the inverse frequency of the documents [33]. The measure suppresses words that are common in the corpus and emphasises rare terms that are characteristic of a given document, making it effective for thematic differentiation and calculating similarity between documents [33]. The result is a sparse "document-term" matrix with TF-IDF weights, which often outperforms simple counting representations in information retrieval and text classification tasks.
- **Embeddings** represent an advanced approach to numerical text representation, in which dense vector representations are trained to capture the semantic and contextual dependencies between words [34], [35]. Unlike classical sparse models, static embeddings assign a fixed vector to each word, while contextual models based on transformer architectures (e.g., BERT and its derivatives) generate representations that depend on the specific use of the word [35], [36]. Although they require significant computational resources and are more difficult to interpret, embedding methods are

becoming the standard in modern natural language processing due to their high semantic expressiveness and efficiency in analysing and comparing texts [37].

Popular proximity metrics that are used are cosine proximity and Jaccard coefficient, due to their simplicity and efficiency [38].

CHAPTER 3. MODELS AND SOFTWARE COMPONENTS FOR PERSONALIZED CONTENT PRESENTATION IN DIGITAL LIBRARIES

This chapter presents the methodological basis and architectural organisation of the proposed solution for presenting personalised content in the form of texts with similar content and personalised recommendations in a digital library. Personalization is understood as the integration of content-based, behavioural, and semantic indicators through which the system adapts the presentation of information resources to the usage context and the established preferences of a specific user, while simultaneously providing transparent grounds for the recommendations.

The approach is **hybrid** and focuses on **text resources** such as **multi-topic periodicals in Bulgarian**. It integrates the processing of information resources, logs of user interactions with the system, and enriched metadata into a unified architecture, in which computationally intensive steps are performed as separate processes outside the operational flow in an asynchronous layer. It includes two main processes: the semantic representation of texts by constructing a **similarity matrix** between documents, and analysis of user interactions by constructing a **sparse "user-document" matrix** and a **popularity vector**. The data is processed incrementally and periodically in batch mode, which allows for efficient updating of only newly received information and ensures scalability for large collections.

The actual personalisation, or more precisely the **interactive layer**, is based on pre-calculated operational structures, which guarantee low latency, consistency of results, and reproducibility. It implements two types of functionalities: displaying **"similar documents"**, based on the similarity matrix, and **generating personalised recommendations** by combining individual preferences and global popularity. The model aims to overcome limitations such as "cold start", dependence on large amounts of behavioural data, and high computational complexity.

3.1. Conceptual model and architectural framework for personalized content presentation in a digital library

The proposed hybrid solution is based on a multi-layer architecture in which computationally intensive operations are performed in a pre-processing stage (asynchronous layer), while the interactive layer uses pre-computed structures and aggregated metrics. This organisation aims to ensure scalability by moving heavy computations off the critical path.

During the preliminary preparation stage, **text resources** and **user interaction logs** are used, supplemented by **extracted named entities**, to construct compact and incrementally updatable structures outside of direct user interaction flow. At this stage, the text corpus undergoes unified pre-processing and is converted into compact vector representations. A document similarity matrix is built on top of them, summarising both global thematic proximity and local matches, optionally enriched with information from extracted named entities. This matrix serves both as the basis for the "similar documents" functionality and as the core of the recommendation algorithm.

An analysis of user interaction logs is performed, from which a sparse "user-document" matrix and a global popularity indicator are constructed. The hybrid algorithm combines content proximity, individual history, and popularity, providing adequate recommendations even with limited user data.

The interactive layer works entirely on these pre-calculated structures and implements two main modes: "similar documents", based on semantic proximity between texts, and "personalised recommendations," which combine content proximity, individual user history, and global popularity. This ensures low latency, stability as the volume of data grows, and explainability of results.

3.2. Service for extracting and structuring named entities

A key element of the architecture is the named entity extraction service, which enriches the presentation of documents with additional structured semantic information and is used to determine the degree of similarity between multi-topic documents. Its main function is to *enrich the documents' metadata* through structured references to names, organisations, geographical places, and other significant objects, which build on standard vector representations and *provide additional semantic context*. In this way, the service addresses key limitations identified in the literature review by building an assessment that is not based on a single component but combines several complementary sources of information. This contributes to more stable results, mitigates the "cold start" effect, increases explainability, and provides scalable semantic enrichment of metadata.

The named entity extraction service is implemented as a standalone **asynchronous** service in batch mode, which processes documents in batches with minimal pre-processing so as not to interfere with entity recognition. The extracted named entities are stored as enriched metadata and support search, similarity assessment, and recommendation generation. In this

way, the service acts as a connecting component between text content, metadata and personalisation algorithm.

3.3. Similarity matrix and the method of multi-component similarity assessment

This subsection presents in detail the basic operational structure for modelling semantic proximity in the corpus - the **document similarity matrix** - and justifies the method of multi-component measure. The input information for generating the matrix comes from two complementary sources: (1) the **texts**, and (2) the **named entities** extracted from them. A multi-component measure is defined on this basis, and the result is organised as a similarity matrix. The structure constructed in this way ensures both explainability and operational efficiency and serves as a basis for both navigation through "similar documents" and the generation of personalised recommendations. To create the similarity matrix, the data goes through several stages:

- **Data pre-processing.** The data pre-processing stage aims to provide a reliable basis for similarity calculation by reducing noise and normalising the text. The process includes cleaning up irrelevant characters, letter register standardisation, lemmatisation to reduce sparsity, and stop word removal. Optionally, synonym enrichment is applied to limit lexical variations and improve the stability of proximity measures.
- **Data vectorization.** The purpose of vectorization is to transform pre-processed texts into numerical representations suitable for artificial intelligence and machine learning algorithms. Classical vectorization methods such as bag-of-words and TF-IDF provide transparent but shallow representations, as they do not capture semantic similarity between words. This limitation is addressed through embeddings, which represent words as vectors in a multidimensional space and capture semantic and contextual relationships; contextual models, in particular, offer higher accuracy in similarity and recommendation tasks. For vectorization, an approach based on contextual embeddings derived from a transformer architecture supporting the Bulgarian language was employed, providing a balance between representation quality and computational efficiency. Documents are segmented into overlapping fragments followed by aggregation at the document level, which preserves thematic structure and ensures stable and reproducible representations under incremental processing.
- **Dimensionality reduction in textual data.** Documents in digital libraries are often lengthy, multi-thematic, and characterized by substantial semantic overlap, which

complicates similarity measurement and increases computational complexity. A common solution is the application of dimensionality reduction techniques aimed at reducing the number of features while preserving essential semantic information, thereby limiting noise and redundant correlations and facilitating interpretability [39]. Despite the advantages of linear and nonlinear methods such as PCA, NMF, and UMAP, their applicability in dynamic environments is limited, as projection models depend on the initial data distribution and require full recomputation when thematic changes occur in the corpus. *For this reason, the dissertation adopts an alternative approach in which, instead of applying additional mathematical reduction, an encoding model is employed that inherently generates compact vector representations.* The selected MiniLM model produces dense 384-dimensional vectors, which *significantly reduce memory requirements compared to classical sparse representations, without sacrificing semantic discriminative capacity.* This approach enables linear and incremental addition of new documents without the need to recompute indexes, making it more suitable for real-world digital library environments.

A set of **complementary components** for assessing the similarity between documents i and j is defined on the vectorised representations of the texts, aiming to cover different aspects of semantic proximity:

- **Global similarity** ($S_{mean}(i, j)$) provides a generalised and stable measure of *thematic similarity*. The similarity between documents is assessed using a cosine similarity applied to their vector representations, with a higher value reflecting a greater degree of content proximity.
- **Local similarity** ($S_{best-match}(i, j)$) aims to capture *strong partial matches* of content at the fragment/segment level. This approach is particularly relevant for digital periodicals, where a single document combines heterogeneous materials (sections, articles, news items) on different topics. The measure favours pairs of documents that share clearly distinguishable thematic blocks, even when their overall themes differ.
- **Topic similarity** ($S_{topic}(i, j)$) introduces a coarser but interpretable structure by grouping content into broad topics. In this approach, all fragment vectors are clustered using a fuzzy clustering algorithm (fuzzy c-means) into C thematic centroids. Each fragment is assigned a degree of membership to the respective topics.

- **Named entities similarity** ($S_{named\ entities}(i, j)$) adds a structured *semantic signal*, which is particularly informative in periodicals. For each document i , a set E_i is formed, where E_i is the set of named entities. Metadata similarity is specified by the **Jaccard** coefficient – this component provides additional proximity to documents that share significant common objects.

The linear combination of these components forms a balanced, accurate, and explainable assessment of similarity, with each contributing a different perspective to semantic proximity. Each component is normalised to a comparable scale, weighted by a set of coefficients, and summed. This forms a compact, sparse representation consistent with the global index of the documents. The individual component scores, calculated on the vector representations of the texts, are combined into the following multi-component measure reflecting semantic proximity:

$$S_{semantic}(i, j) = \alpha * S_{mean}(i, j) + \beta * S_{best-match}(i, j) + \gamma * S_{topic}(i, j), \alpha + \beta + \gamma = 1$$

The final score is formed by adding the component of named entities:

$$S_{final}(i, j) = (1 - \lambda) * S_{semantic}(i, j) + \lambda S_{named\ entities}(i, j). \quad 0 \leq \lambda \leq 1$$

Based on this multi-component measure, which includes semantic similarity, local matches, thematic profiles, and named entities, the final similarity scores between the documents are calculated, and the resulting matrix is stored in the form of a sparse **k-nearest neighbours (k-NN) index**, which limits the memory used and speeds up access [25]. In this form, it performs a dual role: direct input to the "similar texts" module and the core for the subsequent generation of "personalised recommendations".

The hypothesis to be validated in the following chapter is that, for the current corpus of information resources - multi-thematic periodical publications - this multi-component measure will be able to better capture the similarities and differences among the resources.

3.4. User-document matrix and implicit assessment

The logs of interactions between users and the system represent the second main source of information in the proposed architecture, alongside the text resources. The purpose of processing them is to obtain a reliable and compact representation of actual behaviour, which can be used as an implicit assessment of interest and serve as input to the hybrid model for generating personalised recommendations.

As with the similarity matrix, in order to create the **"user-document" matrix**, access logs go through several pre-processing steps:

- **Filtering and cleaning the logs.** The first step in pre-processing user interaction logs involves cleaning irrelevant records – since only "access resource" events are used, administrative operations (creation, modification, deletion of objects), system events, etc. are filtered. After the filtering, each record is reduced to a minimum set of fields: user identifier (u), document identifier (i), timestamp, and action type.
- **Aggregation of implicit ratings.** For each user u and document i , the number of views $c_{u,i}$ (requests of type "access resource") is recorded. Due to the lack of explicit ratings, those logs are interpreted as an implicit indicator of interest and are transformed into an implicit rating $w_{u,i}$ through a monotonically increasing but saturating function. Thus, a single view sets a base weight; additional views increase it with decreasing increments. On the basis of these evaluations, a sparse interaction matrix is constructed.
- **Construction of a global popularity vector for documents.** In parallel with individual interactions, a global popularity vector for documents is constructed based on the number of views, transformed again by a monotonically increasing, saturating function, and normalised in the interval $[0,1]$. This indicator is used as an additional component in the recommendation model, especially in cases of limited individual history, functioning as a stable global signal that mitigates the "cold start" effect without replacing semantic proximity, and is defined by the formula: **popularity**(i) = $\sum_u w_{u,i}$.

3.5. Operational structures and mechanisms for updates

The proposed solution must function in a dynamic environment in which text resources are constantly being added, modified, or deleted, and new interaction logs are accumulating. It is necessary to ensure the consistency of identifiers and pre-calculated operational structures (similarity matrix, user-document matrix, and global popularity vector) as well as their reliable storage and procedures for periodic/incremental update, so that the system maintains accuracy, stability, and performance with increasing data volume and load.

For this reason, the two main flows of information in the system – content and behaviour – are coordinated through a shared index of persistent identifiers, which ensures unambiguous representation of documents in all operational structures. *In this way, the semantic proximity between the texts and the observed user behaviour can be consistently combined within a common reference model, without the risk of inconsistencies between the different representations.*

The system records all types of interactions, not only for resource access, but also events such as resource creation, deletion, and modification, which are used to **maintain the relevance and consistency of representations**. Incremental updating is achieved by processing only new records, which allows documents to be added, edited, or deleted through local updates to the affected structures, *without the need for a complete recalculation*.

As the volume of data increases, a combination of gradual addition and less frequent full recalculation is applied when necessary to restore maximum consistency. Computational-heavy operations are performed outside of user request handling, with the interactive layer working solely with pre-computed matrices and indexes. This ensures short and predictable response times, reproducibility of results, and scalability with a dynamically changing corpus and growing number of users.

3.6. Modules for recommending personalised content

The modules for personalised recommendations in digital libraries are presented, which combine semantic proximity between documents, observed user behaviour, and global popularity in a unified, parameterizable framework. The model aims to provide relevant and explainable recommendations even with a limited individual history, while maintaining operational efficiency and scalability through the use of pre-computed structures.

When processing data to create operational structures, the data processing sequence allows for parallel execution, but it is methodologically appropriate *to first extract named entities through a separate service so that the subsequent similarity matrix can also include this additional semantic contribution. The system interaction logs can be processed in parallel with the creation of the similarity matrix to create a sparse "user-document" matrix and a global popularity vector*.

A shared index for the documents is used for both the similarity matrix and the user-document matrix, which ensures a consistent transition between navigation through "similar documents" and personalised recommendations.

The **"similar documents" functional module** is user-invariant and is based entirely on a similarity matrix constructed based on the text resources, which combines global and local semantic proximity, as well as additional contributions from named entities. Extraction is simple and fast: the system locates the active document by identifier, extracts the corresponding row/list of neighbours, excludes the document itself, and applies a threshold and/or k-restriction to return the most relevant results. Since pre-calculated structures are used, the response time

is short, and explainability is high - the rationale for the suggestions can be traced through shared topics, matching fragments and/or common named entities.

The **functional module for generating personalised recommendations** integrates content proximity with the user's individual history and a global popularity indicator. For **users with sufficient accumulated interactions**, recommendations are formed by transferring interests to semantically similar documents. For **new or anonymous users**, the popularity indicator is primarily relied upon, and as interactions accumulate, the same user smoothly transitions to full hybrid mode without changing the architecture. For **less active users**, the hybrid measure ensures stability of results: the content indicator remains dominant and compensates for sparse behavioural links, returning semantically similar suggestions even with a small history. For **completely new documents** ("cold start" for elements) without accumulated interactions, the same content component prevents their isolation - they are included in the recommendations of readers whose previous texts are semantically close to the new document, with their contribution being adjusted with a smaller score until sufficient behavioural information appears. The relevance score of personalised recommendations is set with the following formula:

$$score(u, d) = \sum_{i \in history(u)} w_{u,i} * S_{final}(i, d) + \partial(u) * popularity(d)$$

The hypothesis to be tested in the following chapter is that this hybrid measure generates higher-quality personalized recommendations-particularly in edge case scenarios such as the cold-start problem, sparse user history, and topic shift - compared to collaborative item-based or content-based filtering approaches.

3.7. Explainability and ethical principles in the selection of personalised content

In the context of personalisation in digital libraries, transparency, explainability, and ethical data management are becoming increasingly important, often taking a back seat to the accuracy of recommendations [40], [41], [42]. This identified gap justifies the need for these aspects to be integrated at the architectural level when designing personalisation systems.

In the proposed architecture, explainability is embedded in the very design of the hybrid model through a decomposable multi-component evaluation function that clearly separates the contribution of semantic proximity, thematic overlap, and named entities. This allows the generation of explicit and intuitively understandable explanations for each recommendation, without the use of external interpretation models [40], [41], [43]. For better comprehensibility,

numerical scores are transformed into linguistic labels, calibrated empirically, and in line with research on the cognitive accessibility of XAI approaches [44], [45].

A fallback strategy using globally popular documents prevents the amplification of biases in sparse data [46], as well as through saturation functions that prevent the domination of single repeated interactions [42]. At the same time, principles for data minimisation and protection are applied through limited scope registers, pseudonymisation, and targeted use of extracted named entities, and traceable access control, in line with best practices for explainable and accountable systems [44], [47], [48], [49], [50].

CHAPTER 4. EXPERIMENTAL IMPLEMENTATION AND ANALYSIS OF PERFORMANCE TESTING

This chapter presents the implementation of a model for personalised content presentation in a digital library, developed in accordance with the architecture described in Chapter 3. Personalisation is based on a combined analysis of the content similarity between documents, user behaviour, and additional semantic information from extracted named entities. The main architectural components are described, including a module that creates the similarity matrix, a module that calculates the user-document matrix and global popularity vector, a named entity extraction service, and functional modules for providing "similar documents" and personalised recommendations. The following subsections discuss the data structures and algorithms used and the experimental verification of the model on synthetic and real data.

4.1. Building a technological environment, test data and protocol for experimental verification

The software implementation of the proposed model is based on a modular architecture optimized for high-performance matrix calculations and text processing in Bulgarian. The technological stack was selected with a view to effective work with large corpora, reproducibility of results, and integration between asynchronous computing modules and the interactive layer. The main programming language is Python 3.13+, due to its wide application in machine learning and the availability of specialised libraries.

Sentence-transformers (MiniLM) [51] are used to implement the algorithmic components, to generate semantic vector representations, scikit-learn [52] was used for pre-processing and similarity metrics, and scikit-fuzzy [53] for thematic modelling using Fuzzy C-Means. Linguistic normalisation of Bulgarian text is performed using simplemma [54] and a set of stop words in Bulgarian [55].

The computational core of the system is based on NumPy [56] and SciPy [57] for linear algebra and sparse structures (CSR), including the calculation of the Jaccard coefficient on sets of named entities, as well as PyTorch [58] for tensor calculations of Transformer models. The implementation supports hardware acceleration via CUDA and MPS, with semantic vectors stored in a semi-precision format (Float16) to reduce memory footprint while maintaining accuracy.

The empirical evaluation is based on a corpus of 1,000 text-based multi-thematic resources from the digital library of the Ivan Vazov National Library in Plovdiv [59], as well

as synthetic sets for validating the multi-component similarity measure and the personalised recommendation module. Synthetic user profiles were also created, covering basic scenarios such as "cold start", dynamic change of interests, and limited history.

4.2. System architecture

The architecture follows the principle of a clear separation between computationally intensive processes and a lightweight interactive part for handling requests. It is structured in two layers – an **asynchronous** and an **interactive layer** (see Fig. 16).

The asynchronous layer combines all computationally intensive processes through which the text corpus and interaction logs are converted into pre-computed representations and indexes used by the interactive part. This separation ensures low latency, predictable behaviour, and sustainable scaling with increasing data volumes and user activity.

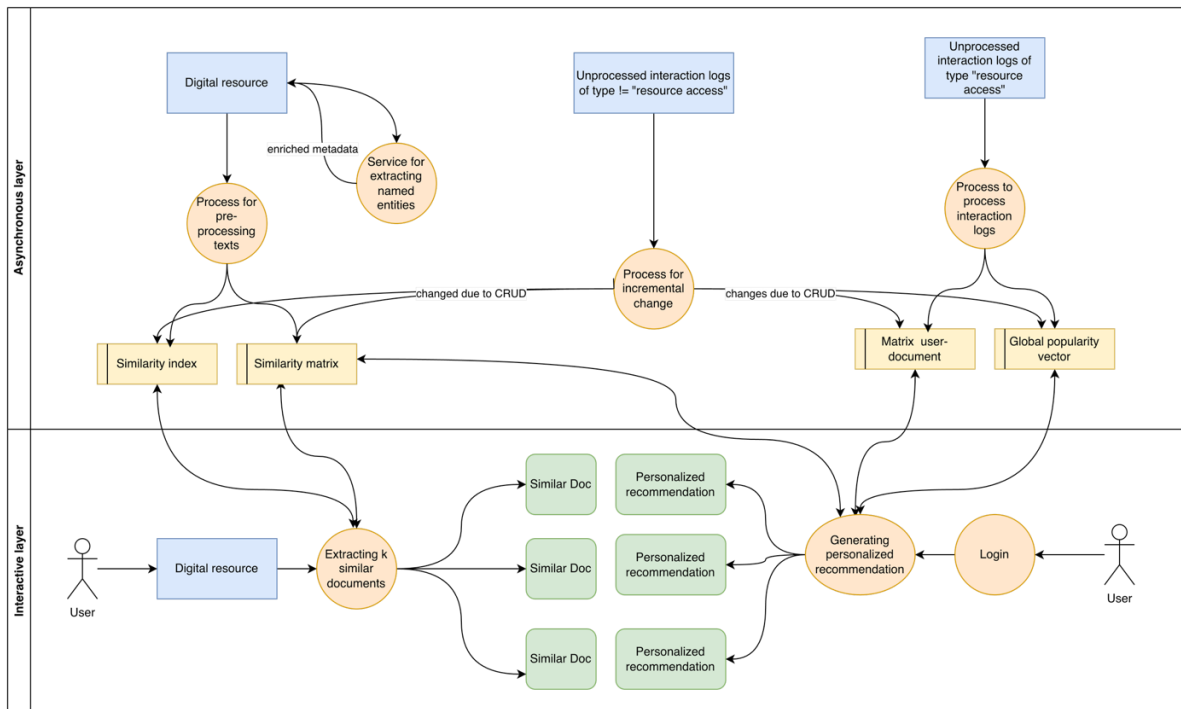


Figure 16. Architecture of the modules for providing personalised content

Three independent components operate in the asynchronous layer, synchronised through shared identifiers: a named entity extraction service that enriches documents with structured semantic information; a similarity matrix generation module that builds a sparse index of the similar documents, based on semantic and thematic features; and an interaction processing module that constructs a sparse "user-document" matrix and a global popularity vector.

Once the asynchronous layer creates and updates all the necessary representations and indexes, the interactive layer implements direct service of user requests, working entirely on these pre-calculated structures. The interactive layer is implemented through two functional modules. The "similar documents" module extracts the resources that are most similar in content to the current document from the similarity matrix. The "personalised recommendations" module combines information about the content similarity between documents, individual user history and global popularity indicators, excluding resources that have already been viewed. In the absence of sufficient user history, representative popular resources are used.

4.3. Service for extracting and structuring named entities

Within the architecture, a separate named entity extraction service was implemented, which aims to enrich document metadata with structured context (e.g. personal names, organisations, geographical objects).

The extracted lists of named entities are used in two main ways: as an additional semantic indicator when building the similarity matrix between documents and as a search aid. In this way, named entities simultaneously increase the precision and explainability of recommendations, as well as navigation within collections. The service operates asynchronously in batch mode, outside the critical path of user requests.

The texts are processed with minimal transformations, as aggressive cleaning would impair the quality of named entity recognition. The approach is based on pre-trained transformer models selected based on good coverage of the Bulgarian language. A selection and experimental comparison were conducted on a sample of documents, taking into account both accuracy and processing time.

The named entity extraction algorithm segments long documents into overlapping fragments, from which entities are sequentially extracted and subsequently merged. Empirical analysis shows that a segmentation size of about 200 tokens achieves an optimal balance between computational efficiency and sufficient context, while longer inputs lead to a decline in performance. In order to increase robustness and efficiency, a dual-model approach is applied, combining the results of the two models with the best accuracy-speed ratio.

After extraction, post-processing is performed, including register normalisation, merging the results of the two models by frequency and confidence maximums, and filtering with configurable thresholds for the number of occurrences and confidence. The resulting

entities are stored as enriched metadata, aligned with document identifiers, and ready for direct integration into similarity measures and search.

In summary, the named entity extraction service provides sustainable and scalable enrichment of documents with structured semantic context. The combination of segmentation, empirically selected fragment size, dual-modal analysis, and targeted post-processing ensures high extraction quality, low latency, and good integration with the other components of the system.

4.4. Similarity matrix and multi-component assessment method. Functional module for selecting “similar documents”

In order to increase resistance to lexical variations, a synonym dictionary for the Bulgarian language was used, extracted from Open Multilingual WordNet [60]. Synonym enrichment is applied only in the pre-processing stage of the "similar documents" functionality, by adding a limited number of representative synonyms before text segmentation. This reduces sensitivity to superficial lexical differences without affecting the interactive layer and user explanations.

The module for calculating similarity between text documents is the core of both the "similar documents" functionality and the personalised recommendations. It operates in the asynchronous layer and is designed with a focus on interpretability, traceability, and computational efficiency, aiming to build a reliable and scalable representation of semantic similarity. The input includes a normalized text corpus and extracted named entities for each document. The output is a sparse symmetric similarity matrix, presented as a list of top-k neighbours and an accompanying dictionary for shared identifiers.

The process of generating the matrix begins with loading the text corpus and the shared index of document identifiers, followed by unified text preparation – normalisation, lemmatisation, stop word filtering, and, optionally, synonym enrichment. Long documents are segmented into overlapping fragments, taking into account the maximum allowable length of the input for the language model, with the overlap preserving semantic continuity and preventing information loss at the boundaries.

Each fragment is encoded into a 384-dimensional vector using the multilingual MiniLM model, after which aggregated representations in the form of averaged vectors are derived at the document level. When the thematic layer is activated, membership profiles obtained through fuzzy clustering over the segment vectors are also computed. In parallel, the extracted named

entities are loaded, and a structured representation is constructed, which serves as an additional input for similarity assessment.

The final similarity score between two documents is defined in the interval $[0, 1]$ and is obtained through a multi-component linear combination of the content component, which includes global similarity between the averaged vectors, local similarity through maximum match between segments, and topic similarity based on the probability profiles from fuzzy grouping. This semantic layer is complemented by a factual indicator based on the overlap of named entities.

The resulting similarity matrix is materialized into compact, sparse structures optimized for fast access and is stored together with a dictionary enabling bidirectional mapping between document identifiers and matrix indices. This matrix is used by the “similar documents” functional module in the interactive layer, which is reduced to direct access, filtering, and value-based ranking.

The quality of the similarity matrix and the weighting coefficients $(\alpha, \beta, \gamma, \lambda)$ of the individual components are calibrated through a multi-stage experimental procedure, combining parametric optimisation, ablation analysis, and qualitative assessment on controlled samples of increasing volume.

For validation purposes, a "gold standard" has been constructed that combines lexical overlap and named entity overlap. Validation is performed on small, medium and control samples (200, 500 and 1000 documents), with parameter optimisation, similarity matrix construction and independent checks performed for each sample. The analysis includes structural verification of the matrix, measurement of lexical proximity, and verification of semantic connectivity through shared named entities, which is particularly indicative of thematic proximity in periodicals.

The experimental results show that the optimal value for the weight of named entities is $\lambda=0.5$, which defines a symmetric hybrid between the semantic and factual layers. This is justified by the specificity of periodicals, in which the information value is highly concentrated around specific individuals, organisations and locations. Calibration of the remaining parameters leads to $\alpha = 0.47$, $\beta = 0.07$ and $\gamma = 0.47$, which shows the leading role of global semantics and thematic modelling and the secondary importance of local matches. This suppresses the noise from repetitive columns, advertisements, or standard passages characteristic of periodicals.

Ablation analysis reveals consistent dynamics: the basic semantic model provides broad coverage, the addition of a thematic layer increases selectivity, and the inclusion of named entities restores valid links through a clear factual context. The final multi-component measure achieves a balance between precision, robustness, and explainability, making it suitable for multi-thematic digital libraries. Additional experimental verification with a synthetic corpus confirms that the full multi-component measure outperforms the baseline configurations in terms of internal cohesion, differentiation of unrelated topics, and capture of partial links.

4.5. Sparse “user-document” matrix, hybrid algorithm, and functional module for generating “personalized recommendations”

This subsection validates the behaviour of the personalised recommendation module, based on item-based collaborative filtering, in key and edge-case scenarios. To isolate random influences, a synthetic set of documents organised into three thematic clusters (A, B and isolated C) is used, with strong co-occurrence between A and B, while cluster C remains without behavioural links. Additionally, a globally popular document was included, used as a fallback strategy in the absence of history.

The verification covers the following scenarios:

1. a cold start for a user, where the system correctly switches to non-personalised recommendations based on global popularity;
2. exhausted isolated cluster, where a backup strategy is activated in the absence of new relevant items;
3. change of interest, where single interactions outside the dominant cluster serve as a signal to switch to a new topic;
4. a mixed profile with a balanced history, leading to proportionally represented recommendations from different clusters without "leaking" into isolated topics;
5. a cold start for an element, where the content component allows for the relevant inclusion of a new resource, albeit with a lower rating;
6. a sparse behavioural history, where the content indicator compensates for weak co-occurrence links.

The verification combines a qualitative check of the logic of the recommendations and a quantitative analysis of the distribution of results in the first k positions. The results show that the module generates predictable and accurate personalised recommendations and remains

useful in the absence or scarcity of data and responds consistently to dynamic changes in user interests.

4.6. Limitations and validity of the proposed architecture

The proposed architecture and hybrid recommendation model have been developed with a view to the specifics of the available corpus, the type of log records, and the text processing methods used. For a correct interpretation of the results, it is necessary to clearly distinguish the main limitations and scope of validity of the model. From a data perspective, the model is most reliable for multi-topic periodicals in Bulgarian, and its transfer to other languages or subject domains requires recalibration and additional empirical evaluation. The quality of the text content is a significant factor – OCR errors, missing segments, etc. introduce noise that directly affects vector representations, similarities, and extracted named entities and cannot be completely eliminated.

The content model is limited by its dependence on pre-trained language models, which are not optimised for all domains and may miss rare, obsolete or specific terms. The use of named entities as an additional indicator depends on the reliability of the recognition module, as errors or omissions can affect similarity scores, although filtering by frequency and confidence mitigates this effect. Multi-component measure requires the selection of thresholds and weighting coefficients, which are determined empirically and influence the final similarity structure.

Behavioural data is implicit in nature: each view is interpreted as a positive signal, with no guarantee of actual reading or satisfaction. This introduces uncertainty, which is compensated for by combining it with content metrics and popularity, but cannot be completely eliminated. The distribution of interactions is highly uneven, with some documents having a sparse history and anonymous users being treated in aggregate.

Technical limitations arise from the computational cost of vectorisation, similarity matrix construction, and thematic modelling, which necessitates batch processing, limiting stored similarity features, and careful selection of update frequency. The use of large language models for named entity extraction increases hardware requirements and may require specialised infrastructure. Furthermore, recommendations do not reflect every new interaction in real time due to the batch processing mode - a trade-off between timeliness and computational feasibility.

In terms of validity, internal validity is supported by clearly defined and repeatable processes for data preparation and operational structure creation, which ensure reproducibility of results. External validity is limited to environments with a similar content structure and interaction logs and requires parameter adaptation and re-evaluation when applied in a different context. This section outlines the limits within which the proposed architecture is correct and useful and serves as a basis for a realistic interpretation of the results and planning of future improvements.

CONTRIBUTIONS OF THE DISSERTATION

Scientific Contributions:

1. A conceptual model and architectural framework for personalized content presentation in a digital library have been developed. These include an asynchronous layer for building operational structures (similarity matrix, “user-document” matrix, global document popularity vector, named entity structures) and a low-latency interactive layer. The operational structures and the relationships between them are formalized, ensuring reproducibility, traceability, and compatibility between modules.
2. A method is proposed for a multi-component similarity assessment between multi-topic documents, defined as a linear combination of indicators for global semantic proximity, local fragment matches, topic profiles, and named entities. The similarity scores between documents are stored in a similarity matrix, which simultaneously supports the “similar documents” functionality and aids in the generation of personalized recommendations.
3. A hybrid algorithm has been developed for generating and delivering content relevant to user needs in a digital library. It is element-oriented and aggregates the content similarity between candidate documents and elements from the user’s individual history on a pre-calculated similarity matrix. To ensure stability, a fallback strategy has been introduced, based on a global popularity vector of the documents, which is applied in edge cases such as a “cold start” and sparse history. This ensures the usefulness of the recommendations even with a limited number of observed interactions.

Scientific and Applied Contributions:

1. A service has been implemented to extract and structure named entities from Bulgarian-language texts as an additional information indicator, which enriches the descriptive

data of documents, improves similarity scoring, and supports searching by structured fields.

2. A functional module for selecting “similar documents” has been implemented, which generates personalized content based on a configurable multi-component similarity assessment between documents. The module ensures retrieval of the closest neighbors with low latency, incremental updates without recalculating the entire operational structures, and full compatibility with the system’s interactive layer.
3. An element-oriented hybrid algorithm has been implemented, based on a sparse “user-document” matrix. It forms the core of a functional module for personalized recommendations. It is used to generate recommendations for “similar documents” closely related to the user’s current interaction history. Candidate documents are formed from nearby neighbors, previously viewed resources are excluded, and they are ranked by combining content contribution and implicit weights. If necessary, the global popularity of documents is used in cases of a “cold start” or when personalized recommendations are exhausted.
4. Mechanisms have been defined and implemented for the periodic updating and incremental expansion of the operational structures and monitored interactions, ensuring the system’s scalability and resilience in the face of growing data volumes and increasing user activity.
5. An experimental validation of the “similar documents” functional module has been performed. A systematic review of the parameters and an ablation analysis of the multi-component evaluation (semantics, local matches, thematic profiles, named entities) were conducted to calibrate the weights and assess the individual contribution of each component to the final result.
6. An experimental validation of the algorithm and the functional module for generating personalized recommendations was performed. Controlled scenarios covering key edge cases (cold start for a user and for an item, sparse data, changing and mixing interests, and recommendation exhaustion) were executed. It has been empirically proven that the hybrid aggregation of content similarity, user history, and popularity leads to predictable and stable behavior of the recommendation module, including in boundary scenarios.

APPROBATION

Some of the results presented in this dissertation research have been achieved and validated with the author's participation in scientific projects and programmes:

- National Scientific Programme "Young Scientists and Postdoctoral Researchers – 2", approved by Council of Ministers Decision No. 206 / 07.04.2022.
- National Interdisciplinary Research E-Infrastructure for Resources and Technologies for Bulgarian Language and Cultural Heritage, integrated within the European infrastructures CLARIN and DARIAH (CLaDa-BG), Ministry of Education and Science, programme "National Roadmap for Scientific Infrastructure", 2018-2027, contractors: ICT-BAS (coordinator), IMI-BAS (partner), DOI-167/28.07.2022, DOI-324/01.12.2023, DOI-97/26.06.2025

List of publications on the topic of the dissertation

1. **Mitreva, E.**, Paneva-Marinova, D., Georgiev, V., Nikolova, A., Pavlov, R. A hybrid approach for personalized and intelligent content recommendation in digital libraries. *Applied Sciences*, Vol. 16, No. 6, Article 2756, MDPI, 2026, ISSN 2076-3417, DOI: <https://doi.org/10.3390/app16062756>, SJR (Scopus): 0.521, Q2 (Web of Science), indexed in Scopus and Web of Science.
2. **Mitreva, E.**, Paneva-Marinova, D., Georgiev, V., Nikolova, A.. A Multi-component Similarity Measure for Personalized Content Discovery in Periodical Digital Library Collections. In: Arai, K., Lorenz, P. (eds) *Proceedings of the Computer Vision Conference (CVC) 2026, Volume 2. CVC 2026. Lecture Notes in Networks and Systems*, vol. 1975, Springer, Cham, 2026, ISBN:978-3-032-26210-3, ISSN:2367-3370, DOI: https://doi.org/10.1007/978-3-032-26211-0_22, 357-371. SJR (Scopus):0.165, Q4 (Scopus) (в процес на индексване в Scopus)
3. **Mitreva, E.** Improving short text classification with semi-supervised learning. *TEM Journal*, Vol. 15, No. 1, UIKTEN - Association for Information Communication Technology Education and Science, 2026, pp. 876–883, ISSN 2217-8309, DOI: <https://doi.org/10.18421/TEM151-80>, SJR (Scopus): 0.242, Q4 (Web of Science), indexed in Scopus and Web of Science.
4. **Mitreva, E.**, Georgiev, V., Nikolova, A. Classification of short noisy text. In: *Proceedings of the International Conference on Computer Systems and Technologies 2024 (CompSysTech '24)*, ACM International Conference Proceedings Series, ACM,

New York, USA, 2024, pp. 227–231, ISBN 979-8-4007-1684-3/24/06, DOI: <https://doi.org/10.1145/3674912.3674935>, SJR (Scopus): 0.253, indexed in Scopus.

5. **Mitreva, E.**, Nikolova, A., Georgiev, V., Gigova, A. Personalization approaches for cultural heritage study. In: *Proceedings of the Digital Presentation and Preservation of Cultural and Scientific Heritage*, Vol. 13, Institute of Mathematics and Informatics – BAS, 2023, pp. 181–188, ISSN 1314-4006, DOI: <https://doi.org/10.55630/dipp.2023.13.17>, indexed in Scopus and Web of Science.

List of reported results

1. Scientific paper presented at the international conference Computer Vision Conference 2026, 21-22 May 2026, Amsterdam, Netherlands, Topic: A Multi-Component Similarity Measure for Personalised Content Discovery in Periodical Digital Library Collections, Authors: **E. Mitreva**, D. Paneva-Marinova, V. Georgiev, A. Nikolova, Date: 22.05.2026
2. Scientific paper presented at the international conference Computer Systems and Technologies 2024 (CompSysTech '24), 14–15 June 2024, Ruse, Bulgaria, Topic: Classification of Short Noisy Text, Authors: **E. Mitreva**, V. Georgiev, A. Nikolova, Date: 15.06.2024
3. Scientific paper presented at the international conference Digital Presentation and Preservation of Cultural and Scientific Heritage, 7–10 September 2023, Burgas, Bulgaria, Topic: Personalisation Approaches for Cultural Heritage Study, Authors: **E. Mitreva**, A. Nikolova, V. Georgiev, A. Gigova, Date: 07.09.2024
4. Scientific paper presented at the Annual Reporting Session of the Mathematical Linguistics Section, IMI-BAS on 22 December 2023, Sofia, Bulgaria, Topic of the paper: Approaches to personalisation in the study of cultural heritage, Author: E. Mitreva
5. Scientific report presented at the Annual Reporting Session of the Mathematical Linguistics Section, IMI-BAS on 18 December 2024, Sofia, Bulgaria, Topic of the report: Classification of short noisy texts, Author: **E. Mitreva**
6. Scientific report presented at the Annual Reporting Session of the Mathematical Linguistics Section, IMI-BAS on 10 December 2025, Sofia, Bulgaria, Topic of the

report: Models and methods for providing personalised content in digital libraries,
Author: **E. Mitreva**

List of citations

Total found citations – 4 (excluding self-citations).

Mitreva, E., Paneva-Marinova, D., Georgiev, V., Nikolova, A., Pavlov, R. A hybrid approach for personalized and intelligent content recommendation in digital libraries. *Applied Sciences*, Vol. 16, No. 6, Article 2756, MDPI, 2026, ISSN 2076-3417, DOI: <https://doi.org/10.3390/app16062756>, SJR (Scopus): 0.521, Q2 (Web of Science)

Cited in:

- Sahid, N. Z., Abdullah Sani, M. K. J., Mohamad, A. N., Ahmad Saleh, A., Baba, J., Adriani Salim, T. (2026). AI-enabled quality as a driver of user satisfaction and digital content engagement in Malaysian ubiquitous libraries: An ISSM approach. *Journal of Librarianship and Information Science*. Advance online publication. doi: <https://doi.org/10.1177/09610006261442582>. Journal ISSN 0961-0006 (print); E-ISSN 1741-6477.

Mitreva, E., Nikolova, A., Georgiev, V., & Gigova, A. Personalization approaches for cultural heritage study. *Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings*, 2023, 13, 181–188. Institute of Mathematics and Informatics – BAS. <https://doi.org/10.55630/dipp.2023.13.17>, ISSN: 1314-4006

Cited in:

- Megawati, C. D., Kian, T. P., & Sutawijaya, B. (2026). Integrating Agile Development and Content-Based Filtering for Personalized Digital Cultural Heritage Applications: A Case Study of Sri Ranggah Rajasa Sang Amurwabhumi. *Sinkron: jurnal dan penelitian teknik informatika*, 10(1), 1-14. ISSN: 2541-2019

Mitreva, E., Georgiev, V., & Nikolova, A. Classification of short noisy text. *Proceedings of the International Conference on Computer Systems and Technologies 2024 (CompSysTech '24)*,

Cited in:

- Gonçalves, J. J. O., Lotufo, T., Nze, G. D. A., & de Mendonça, F. L. (2025). Detecção de Prompt Injection em modelos de Linguagem. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E77), 104-116. ISSN: 1646-9895
- Rianto, R., Humanika, E. S., & Untoro, I. H. T. (2026). Enhancing SVM-Based Classification Performance on Indonesian Sentences through TF-IDF and Directional Augmentation. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, 10(1), 22-35. ISSN: 2580-409X

BIBLIOGRAPHY

- [1] Z. Liu and B. Shao, "A systematic review of library services platforms research and research agenda," *Library & Information Science Research*, vol. 46, no. 4, 2024. ISSN 0740-8188, <https://doi.org/10.1016/j.lisr.2024.101325>.
- [2] C. L. Borgman, "Libraries, Digital Libraries, and Data: Forty years, Four Challenges," *portal: Libraries and the Academy*, vol. 25, no. 3, pp. 39-58, 2025. <https://doi.org/10.48550/arXiv.2506.15055>.
- [3] J. C. Shikali and P. S. Muneja, "Access to Library Information Resources by University Students during COVID-19 Pandemic in Africa: A Systematic Literature Review," *arXiv*, 2024. doi: 10.21203/rs.3.rs-4237695/v1.
- [4] M. Goynov, D. Luchev, D. Paneva-Marinova, G. Senka, K. Rangochev, L. Pavlova, R. Pavlov and L. Zlatkov, "CultIS: Web-based Platform for Intelligent Cultural Content Management," *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 14, pp. 19-36, 2024. <https://doi.org/10.55630/dipp.2024.14.1>.
- [5] M. Goynov, D. Luchev, D. Paneva-Marinova, R. Pavlov and K. Rangochev, "Towards Providing Analytical Services in the Web-Based Platform for Intelligent Cultural Content Management CultIS," *TEM Journal*, vol. 14, no. 4, pp. 2946-2952, 2025. doi: 10.18421/TEM144-05.
- [6] C. G. S. and M. Mulimani, "The Impact of Artificial Intelligence on Library and Information Science (LIS) Services," *Social Science Research Network*, vol. 14, no. 5, pp. 50-56, 2024. doi: <http://dx.doi.org/10.2139/ssrn.4856459>.
- [7] D. Paneva-Marinova, M. Goynov and R. Pavlov, "Enhanced and personalized learning experience in digital libraries," in *In the Proceedings of the 10th annual International Conference of Education, Research and Innovation*, 2017. doi: 10.21125/iceri.2017.0595.
- [8] M. Marzuki, S. F. Z. Azero, A. Alia and M. R. A. Kadir, "A Systematic Literature Review of User Behavior and Personalization in Digital Libraries," *International Journal of Research and Innovation in Social Science*, vol. 9, no. 1, pp. 4830-4842, 2025. doi: 10.47772/IJRIS.2025.9010372.
- [9] V. A. Kumar and M. Chidambaram, "Personalization and User Behavior Analysis in Digital Libraries: A Systematic Review," *Academic Research Journal of Science and Technology (ARJST)*, vol. 2, no. 2, pp. 37-43, 2025. doi: 10.63300/arjst0202202505.

- [10] D. Christozov and S. Toleva-Stoimenova, "Big Data Literacy: A New Dimension of Digital Divide, Barriers in Learning via Exploring "Big Data"," in *Strategic Data-Based Wisdom in the Big Data Era*, 2015, pp. 156-171. doi: 10.4018/978-1-4666-8122-4.ch009.
- [11] V. H. Reji, "Traditional Libraries Vs Digital Libraries: A Comparative Analysis," *Academic Research Journal of Science and Technology*, vol. 1, no. 8, 2025. doi: <https://doi.org/10.63300/arjst10906202501>.
- [12] D. Christozov and E. Mitreva, "Trust in learning from big data: the two sides of the same coin," *Information Systems*, vol. 21, no. 1, pp. 147-152, 2020.
- [13] Z. Fayyaz, M. Ebrahiman, D. Nawara, A. Ibrahim and R. Kashef, "Recommendation systems: Algorithms, challenges, metrics, and business opportunities.," *Applied Sciences*, vol. 10, no. 21, 2020. doi: <https://doi.org/10.3390/app10217748>.
- [14] H. Liqiang and L. Quan, "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*, 2019. doi: <https://doi.org/10.1145/3377672.337805>.
- [15] K. Stefanov, P. Boychev, E. Stefanova and A. Georgiev, "Digital Libraries in Teacher Education," in *Fortieth Jubilee Spring Conference of the Union of Bulgarian Mathematicians*, 2011.
- [16] M. Liao and S. S. Sundar, "When e-commerce personalization systems show and tell: Investigating the relative persuasive appeal of content-based versus collaborative filtering," *Journal of Advertising*, vol. 51, no. 2, pp. 256-267, 2022. doi: 10.1080/00913367.2021.1887013.
- [17] J. B. Schafer, D. Frankowski, J. Herlocker and S. Shen, "Collaborative filtering recommender systems. The adaptive web: methods and strategies of web personalization," in *The Adaptive Web. Lecture Notes in Computer Science*, vol. 4321, 2007, pp. 291-324. doi: https://doi.org/10.1007/978-3-540-72079-9_9.
- [18] S. Kapembe and J. G. Quenum, "A Personalised Hybrid Learning Object Recommender System," in *11th International Conference on Management of Digital EcoSystems*, 2019. <https://doi.org/10.1145/3297662.3365810>.
- [19] P. Davidson, F. Buckermann, M. Steininger, A. Krause and A. Hotho, "Semi-unsupervised Learning: An In-depth Parameter Analysis," in *Lecture Notes in Computer Science*, vol. 12873, 2021. https://doi.org/10.1007/978-3-030-87626-5_5.
- [20] R. S. Nurhalizah, R. Ardianto and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *Jurnal Ilmu Komputer dan Informatika*, vol. 4, no. 1, pp. 61 - 72, 2024. doi: 10.54082/jiki.168.
- [21] E. Najjar and A. M. Breesam, "Supervised Machine Learning a Brief Survey of Approaches," *Al-Iraqia Journal of Scientific Engineering Research*, vol. 2, no. 4, p. 71–82, 2023. doi: 10.58564/IJSER.2.4.2023.121.
- [22] X. Zhang, F. Guo, C. Tao , P. Lei, G. Beliakov and J. Wu, "A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 18, no. 4, pp. 2188-2216, 2023. doi: <https://doi.org/10.3390/jtaer18040110>.
- [23] E. Gomedé, "Understanding Multinomial Naive Bayes Classifier," 11 November 2023. [Online]. Available: <https://medium.com/@evertongomedé/understanding-multinomial-naive-bayes-classifier-fdbd41b405bf>. [Accessed 31 December 2023].
- [24] Y. Wang, H. Lin, C. Li, L. She, L. Sun and J. Wang, "Network Autonomous Learning Monitoring System Based on SVM Algorithm," in *10th International Conference on Wireless Communication and Sensor Networks (icWCNS '23)*, 2023. doi: <https://doi.org/10.1145/3585967.3585984>.
- [25] Y. Sun and Q. Liu, "Collaborative filtering recommendation based on K-nearest neighbor and non-negative matrix factorization algorithm," *The Journal of Supercomputing*, vol. 81, no. 79, 2024. doi: <https://doi.org/10.1007/s11227-024-06537-4>.

- [26] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, 2024. doi: <https://doi.org/10.1186/s40537-024-00973-y>.
- [27] D. R. Anamisa, A. Jauhari and F. A. Mufarroha, "K-Nearest Neighbors Method for Recommendation System in Bangkalanâ s Tourism," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 14, no. 1, pp. 33-44, 2023. doi: [10.21512/comtech.v14i1.7993](https://doi.org/10.21512/comtech.v14i1.7993).
- [28] Y. Lu, H. Xin, R. Wang, F. Nie and X. Li, "Scalable Multiple Kernel k-means Clustering," in *31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, New York, 2022. doi: <https://doi.org/10.1145/3511808.3557690>.
- [29] M. B. Ferraro, "Fuzzy k-Means: history and applications," *Econometrics and Statistics*, vol. 30, pp. 110-123, 2024. doi: [10.1016/j.ecosta.2021.11.008](https://doi.org/10.1016/j.ecosta.2021.11.008).
- [30] S. Li, G. Yuan, M. Yang, Y. Shen, C. Li, R. Xu and X. Zhao, "Improving Semi-Supervised Text Classification with Dual Meta-Learning," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1-28, 2024. doi: <https://doi.org/10.1145/3648612>.
- [31] P. Jain, "Basics of CountVectorizer," 24 May 2021. [Online]. Available: <https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c>. [Accessed 28 November 2023].
- [32] K. Ganesan, "HashingVectorizer vs. CountVectorizer," [Online]. Available: <https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/>. [Accessed 29 November 2023].
- [33] B. Roepke, "A Quick Introduction to Bag of Words and TF-IDF," 21 01 2022. [Online]. Available: <https://towardsdatascience.com/a-quick-introduction-to-bag-of-words-and-tf-idf-fbd3ab84ecbf/>.
- [34] R. Patil, S. Boit, V. Gudivada and J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," *IEEE Access*, vol. 11, pp. 36120-36146, 2023. doi: [10.1109/ACCESS.2023.3266377](https://doi.org/10.1109/ACCESS.2023.3266377).
- [35] Z. Nie, Z. Feng, M. Li, C. Zhang, R. Zhang, D. Long and R. Zhang, "When text embedding meets large language model: a comprehensive survey," *arXiv preprint arXiv:2412.09165*, 2024. doi: [10.48550/arXiv.2412.09165](https://doi.org/10.48550/arXiv.2412.09165).
- [36] H. Cao, "Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark," *arXiv:2406.01607*, 2024. doi: [10.13140/RG.2.2.13267.80162](https://doi.org/10.13140/RG.2.2.13267.80162).
- [37] K. Das, Kamlish and F. Abid, "Advancements in Word Embeddings: A Comprehensive Survey and Analysis," *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, vol. 61, no. 3, 2024. doi: [10.53560/PPASA\(61-3\)842](https://doi.org/10.53560/PPASA(61-3)842).
- [38] A. Levy, B. R. Shalom and M. Chalamish, "A guide to similarity measures and their data science applications," *Journal of Big Data*, vol. 12, no. 1, 2025. doi: [10.1186/s40537-025-01227-1](https://doi.org/10.1186/s40537-025-01227-1).
- [39] A. K. Rastogi, S. Taterh and B. S. Kumar, "Dimensionality reduction algorithms in machine learning: a theoretical and experimental comparison," in *International Conference on Recent Advances in Science and Engineering*, 2023. doi: <https://doi.org/10.3390/engproc2023059082>.
- [40] Y. Zhang and X. Chen, "Explainable Recommendation: A Survey and New Perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1-101, 2020. doi: <https://doi.org/10.1561/15000000066>.
- [41] Tintarev, N. and Masthoff, J., "Designing and Evaluating Explanations for Recommender Systems," in *Recommender Systems Handbook*, 2021, pp. 479–510. doi: https://doi.org/10.1007/978-0-387-85820-3_15.
- [42] S. Milano, M. Taddeo and L. Floridi, "Recommender Systems and Their Ethical Challenges," *AI & Society*, vol. 35, p. 957–967, 2020. doi: <https://doi.org/10.1007/s00146-020-00950-y>.
- [43] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608*, 2017. doi: <https://doi.org/10.48550/arXiv.1702.08608>.

- [44] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, p. 1–38, 2019. doi: <https://doi.org/10.1016/j.artint.2018.07.007>.
- [45] M. T. Ribeiro, S. Singh and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv:1602.04938*, 2016. doi: <https://doi.org/10.48550/arXiv.1602.04938>.
- [46] R. K. Merton, "The Matthew Effect in Science," *Science*, vol. 159, no. 3810, p. 56–63, 1968.
- [47] E. U. A. f. N. a. I. S. (ENISA), "Guidelines for SMEs on the Security of Personal Data Processing," 2016.
- [48] E. U. A. f. N. a. I. S. (ENISA), "Handbook on Security of Personal Data Processing," ENISA, 2017.
- [49] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez and F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, 2023. doi: <https://doi.org/10.1016/j.inffus.2023.101805>.
- [50] J. Z. Forde and M. Paganini, "The scientific method in the science of machine learning," *preprint arXiv:1904.10922*, 2019. doi: <https://doi.org/10.48550/arXiv.1904.10922>.
- [51] "Sentence Transformers," [Online]. Available: <https://huggingface.co/sentence-transformers>.
- [52] "Machine Learning in Python," [Online]. Available: <https://scikit-learn.org/stable/>.
- [53] "scikit-fuzzy," [Online]. Available: <https://pypi.org/project/scikit-fuzzy/>.
- [54] A. Barbaresi, "Simplemma: a simple multilingual lemmatizer for Python," [Online]. Available: <https://github.com/adbar/simplemma>.
- [55] "Stopwords Bulgarian (BG)," [Online]. Available: <https://github.com/stopwords-iso/stopwords-bg>.
- [56] "NumPy," [Online]. Available: <https://pypi.org/project/numpy/>.
- [57] "SciPy," [Online]. Available: <https://scipy.org/>.
- [58] "PyTorch," [Online]. Available: <https://pytorch.org/>.
- [59] "Digital library "National library "Ivan Vazov", Plovdiv," [Online]. Available: <https://digital.libplovdiv.com/bg>.
- [60] B. Group. [Online]. Available: <https://github.com/omwn/omw-data/tree/main/wns/bul>.