



**БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА**

Емануела Димитрова Митрева

**МЕТОДИ И АЛГОРИТМИ ЗА
ПЕРСОНАЛИЗАЦИЯ И АДАПТИВНОСТ В
СРЕДИ ЗА УПРАВЛЕНИЕ НА СЪДЪРЖАНИЕ**

Автореферат

на дисертационен труд

за присъждане на образователна и научна степен „Доктор”

по област на висшето образование 4. Природни науки, математика и информатика,
професионално направление 4.6. Информатика и компютърни науки, докторска
програма „Информатика“

Научен ръководител:

проф. д-р Десислава Панева-Маринова

София, 2026

Дисертационният труд е обсъден и насочен за защита на разширено заседание на секция „Математическа лингвистика“ при Института по математика и информатика при Българска академия на науките на 14.09.2026 г.

Дисертационният труд е изложен в **167** страници и съдържа **15** таблици и **26** фигури. Той включва увод, **5** глави, списък на използваната литература от **202** литературни източници, списък на **5** публикации на автора (1, от които самостоятелна), свързани с представения дисертационен труд.

Номерацията на таблиците и фигурите в автореферата следва оригиналната номерация, използвана в дисертационния труд.

Материалите по защитата са на разположение на интересувалите се в Институт по математика и информатика - БАН, ул. „Акад. Г. Бончев“, блок 8, София.

Автор: Емануела Димитрова Митрева

Заглавие: Методи и алгоритми за персонализация и адаптивност в среди
за управление на съдържание

СЪДЪРЖАНИЕ

Глава 1. Обща характеристика на дисертационния труд.....	4
1.1. Актуалност на проблема.....	4
1.2. Обект, предмет, цел и задачи на изследването.....	4
1.3. Структура на дисертационния труд.....	6
Глава 2. Теоретични основи и анализ на съвременни подходи за персонализация в дигитални библиотеки.....	8
2.1. Дигитални библиотеки – същност и развитие.....	8
2.2. Персонализация в дигитални библиотеки.....	9
Глава 3. Модели и софтуерни компоненти за персонализирано представяне на съдържание в дигитални библиотеки.....	16
3.1. Концептуален модел и архитектурна рамка за персонализирано представяне на съдържание в дигитална библиотека.....	17
3.2. Услуга за извличане и структуриране на именувани същности.....	17
3.3. Матрица на сходство и метод на многокомпонентна оценка на сходство.....	18
3.4. Матрица „потребител-документ“ и имплицитни оценки.....	21
3.5. Оперативни структури и механизми за актуализация.....	22
3.6. Модули за генериране на персонализирано съдържание.....	23
3.7. Обяснимост и етични принципи при селекция на персонализирано съдържание.....	25
Глава 4. Експериментално внедряване и анализ на резултатното тестване.....	26
4.1. Изграждане на технологична среда, тестови данни и протокол за експериментална верификация.....	26
4.2. Архитектура на системата.....	27
4.3. Услуга за извличане и структуриране на именувани същности.....	28
4.4. Матрица на сходство и метод на многокомпонентна оценка. Функционален модул за селектиране на „подобни документи“.....	29
4.5. Разредена матрица „потребител-документ“, хибриден алгоритъм и функционален модул за генериране на „персонализирани препоръки“.....	31
4.6. Ограничения и валидност на предложената архитектура.....	32
Приноси на дисертационния труд.....	34
Апробация.....	36
Списък с авторски публикации по темата на дисертацията.....	36
Списък с докладвани резултати.....	37
Списък с цитирания.....	38
Библиография.....	39

ГЛАВА 1. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1.1. Актуалност на проблема

През последните десетилетия цифровата трансформация промени много начина, по който се съхранява и използва научното и културното наследство. Натрупването на големи масиви от електронни ресурси и широкият дистанционен достъп до тях превърнаха цифровите колекции в неразделна част от научната дейност и образованието. В този контекст дигиталните библиотеки заемат особено място като среди, в които се обединяват дългосрочно съхранение, надеждно описание и организирано предоставяне на разнородни по произход и форма информационни ресурси.

С нарастването на обема и разнообразието на съдържанието обаче възниква съществено предизвикателство: стандартните механизми за търсене и навигация все по-трудно помагат на потребителите да откриват документи, които действително съответстват на техните нужди или интереси. Изобилието от ресурси, липсата на експлицитни оценки, различната степен на структурираност често водят до прекалено много резултати и до затруднения при ориентиране в наличните ресурси. Това поставя на преден план необходимостта от подходи, които не само осигуряват достъп до ресурсите, но и го правят по-селективен и по-съобразен с нуждите на конкретния потребител.

В този смисъл персонализираното представяне на съдържание се разглежда като перспективна посока за развитие на дигиталните библиотеки. Съвременните решения в областта на изкуствения интелект позволяват чрез използване на модели за анализ на текстове, структурирани описания и регистри на взаимодействията между потребителите и документите препоръчителните механизми да подпомагат откриването на близки по съдържание ресурси, при спазване на изискванията за прозрачност и защита на личните данни. Особен интерес представляват хибридните решения, които съчетават няколко източника на информация и различни подходи, с цел постигане на по-устойчиво, обяснимо и приложимо в реална среда препоръчване.

1.2. Обект, предмет, цел и задачи на изследването

Обект на дисертационния труд е процесът на адаптиране и персонализиране на съдържанието в дигитални библиотеки чрез използване на методи и техники на изкуствения интелект и машинното обучение. Изследването се фокусира върху начини,

по които дигитални библиотеки могат да анализират потребителското поведение, предпочитания и контекст, за да предоставят динамично съдържание, съобразено с индивидуалните нужди и интереси на всеки потребител.

Дисертационният труд изследва и систематизира съвременни подходи за персонализирано предоставяне на съдържание в дигитални библиотеки на основата на методи на изкуствения интелект и машинното обучение, като очертава основните проблеми и предизвикателства при тяхната реализация. **Основната цел** е разработването на нови модели, методи и средства за персонализирано представяне на съдържание, които обединяват съдържателни характеристики, данни от регистрите за взаимодействия между потребители и документи, както и метаданни. Целта е на потребителя да се предлагат максимално уместни, обясними и съобразени с нуждите му информационни ресурси при запазена мащабируемост и доказана приложимост в реална среда.

Предметът на изследването са подходи, модели и алгоритми за адаптиране на информационните обектите и ресурси в дигиталните библиотеки с цел предоставяне на персонализирано съдържание.

Изследването се основава на хипотезата, че прилагането на подходящи методи на изкуствения интелект и машинното обучение за адаптиране на съдържание в дигиталните библиотеки води до по-висока степен на персонализация, релевантност и ефективност при предоставяне на информация на потребителите.

В съответствие с тази хипотеза и с оглед постигане на целта на дисертационния труд, са формулирани следните основни изследователски **задачи**:

Задача 1. Да се проучат научните постижения и резултати от актуалните изследвания за използване на методи на изкуствения интелект и машинното обучение за предоставяне на персонализирано съдържание в дигитални библиотеки.

Задача 2. Да се изследват възможностите за прилагане на съвременни методи на изкуствения интелект и обработката на естествен език, използващи големи езикови модели, за извличане на именувани същности от текстови ресурси и интегрирането им като структурирани метаданни, с цел подобряване на възможностите за търсене, както и използването им като допълнителен информационен показател при изграждането на хибридни препоръчващи модули.

Задача 3. Да се създаде концептуален модел на функционални модули за препоръчване на съдържание в дигитална библиотека, базиран на съвременни методи на

изкуствения интелект, който да предлага както информационни ресурси, сходни с текущите разглеждания, така и други ресурси, към които потребителят потенциално би проявил интерес. В рамките на модела да се разработят подходи за обработка и използване на потребителски данни, както и подходи за повишаване на прозрачността и обяснимостта на процеса на препоръчване.

Задача 4. Да се разработи и имплементира прототип на предложените функционални модули и да се проведе експериментално тестване за оценка на неговата ефективност и приложимост.

Задача 5. Да се анализират и интерпретират резултатите от проведените експерименти с цел формулиране на изводи относно качеството на препоръките и потенциала на предложения модул.

1.3. Структура на дисертационния труд

Структурата на дисертационния труд е както следва:

Глава 1. Обща постановка на задачата формулира обекта, предмета, основната цел и конкретните задачи, както и очертава контекста, в който се разглежда персонализираното представяне на съдържание в дигиталните библиотеки.

Глава 2. Теоретични основи и анализ на съвременни подходи за персонализация в дигитални библиотеки представя основните понятия, модели и класификации, свързани с персонализацията и алгоритмите за препоръчване на съдържание и извършва аналитично изследване на научни постижения и резултати от актуални изследвания и научни достижения за използване на методи на изкуствения интелект и машинното обучение за предоставяне на персонализирано съдържание в дигитални библиотеки.

Глава 3. Модели и софтуерни компоненти за персонализирано представяне на съдържание в дигитални библиотеки формулира теоретичната рамка на предложената архитектура за персонализирано представяне на съдържание в дигитална библиотека. Представя се концептуалният модел на системата и ролите на основните ѝ компоненти – два функционални модули за генериране на персонализирано съдържание и една отделна услуга за извличане и структуриране на именувани същности. Главата последователно описва етапите на подготовка на данните и създаването на оперативните структури в асинхронния слой. На базата на създадените структури, в интерактивния

слой се дефинират два водещи начина на предоставяне на персонализирано съдържание: (1) откриване на „подобни документи“, инвариантно за всички потребители; и (2) „персонализирани препоръки“, при което персонализацията се постига чрез хибриден подход: съчетават се съдържателна близост до вече разглеждани ресурси с индикатор от глобалната популярност, така че да се балансират индивидуалните предпочитания и устойчивите тенденции при липса на поведенческа история. Изложението разкрива взаимовръзките между модулите и мотивира избора на хибриден подход.

Глава 4. Експериментално внедряване и анализ на резултатното тестване представя практическата имплементация на предложените модули и компоненти, използваните програмни средства и параметри, методиката за оценка и резултатите от експерименталното тестване, чрез които се оценява приложимостта и ефективността на предложените решения в реална дигитална библиотека.

Глава 5. Заключение и бъдещи насоки за развитие обобщава постигнатите резултати от разработването, анализа и експерименталното внедряване на предложените решения, като се потвърждава тяхната ефективност и приложимост в контекста на дигиталните библиотеки, както и очертава възможни направления за бъдещо развитие чрез разширяване на функционалностите, оптимизация на производителността и интеграция с допълнителни стандарти и интелигентни методи за управление и анализ на дигитално съдържание.

ГЛАВА 2. ТЕОРЕТИЧНИ ОСНОВИ И АНАЛИЗ НА СЪВРЕМЕННИ ПОДХОДИ ЗА ПЕРСОНАЛИЗАЦИЯ В ДИГИТАЛНИ БИБЛИОТЕКИ

Бързото развитие на информационните технологии през последните десетилетия доведе до фундаментални промени в начина, по който се създава, съхранява и споделя знание. Експоненциалният растеж на цифровите ресурси и нарастващата достъпност на информацията трансформираха традиционните подходи към управлението на знание и породиха необходимост от нови методи за неговата организация и използване. Тази трансформация обаче поставя пред изследователите и разработчиците нови предизвикателства, свързани с ефективното управление на информационните ресурси и осигуряването на смислен достъп до знание в условията на информационно пренасищане.

Настоящата глава има за цел да представи теоретични основи и съвременни концепции за персонализация на информационно съдържание, като се фокусира върху принципи, методи и алгоритмични подходи, които определят развитието на тази изследователска област. Разглеждат се утвърдени и иновативни решения, базирани на техники от изкуствения интелект и машинното обучение, които позволяват адаптиране на информационната среда спрямо индивидуалните интереси и поведенчески модели на потребителите. В рамките на анализа се проследява еволюцията на подходите за персонализация и тяхната приложимост в различен контекст, като специален акцент се поставя върху изследването на техни реализации в областта на дигиталните библиотеки.

2.1. Дигитални библиотеки – същност и развитие

Дигиталните библиотеки се възприемат не просто като дигитализирани колекции, а като интегрирани системи за управление на знания, които съчетават информационни ресурси, инфраструктура и услуги в динамична среда [1], [2]. И в този смисъл основната цел на дигиталните библиотеки е да осигурят широк, устойчив и равнопоставен достъп до знание, като подкрепят научните и културните процеси в глобален мащаб [2], [3]. Те се стремят не само към дигитализация на съдържание, но и към изграждане на интелигентна информационна екосистема, която обединява данни, услуги и потребители в динамична среда на взаимодействие [2], [4], [5].

Авторите на [6], [7] допълват, че чрез внедряване на изкуствен интелект и адаптивни технологии дигиталните библиотеки постигат по-висока степен на персонализация и достъпност, като улесняват навигацията и повишават ефективността

на изследователската работа. По този начин тяхното значение надхвърля традиционната функция на съхранение и предоставяне на информация – те се превръщат в активен посредник в създаването, откриването и споделянето на знание, който подпомага развитието на иновации, академични практики и културна памет и се явява активна среда за създаване, споделяне и развитие на знание, което е в основата на съвременните научни и образователни процеси [2], [6].

Също така дигиталната среда променя начина, по който потребителите взаимодействат с информационните ресурси, като създава условия за персонален достъп и динамично търсене. Според [8], дигиталните библиотеки вече не са просто хранилища на информация, а „адаптивни платформи“, които анализират поведенческите модели на потребителите с цел предоставяне на по-уместно съдържание и подобро потребителско преживяване. Персонализацията представлява ключов подход за подобряване на ефективността на информационното взаимодействие. Тя включва адаптиране на съдържанието, интерфейса и функционалностите според индивидуалните характеристики, интереси и поведение на потребителя [9]. Потребителят вече е не просто статичен консуматор на информация и ресурси, а активен участник в изграждането на знание.

В резултат, необходимостта от персонализирани решения в дигиталните библиотеки се разглежда като необходима промяна за подобряване на удовлетвореността на потребителите и за ефективното използване на информационните ресурси. Дигиталните библиотеки предлагащи персонализирано съдържание, не само подпомагат откриването на релевантно съдържание, но и формират по-интелигентна и ангажираща среда за учене и изследване, която отразява дигиталното знание – ориентирана към потребителя, контекста и взаимодействието.

2.2. Персонализация в дигитални библиотеки

Ерата на големите данни утвърждава информацията като стратегически ресурс, а способността за извличане на знание от големи и хетерогенни масиви данни се превръща в ключово конкурентно предимство [10]. Нарастващият обем и сложност на данните обаче затрудняват анализа и ориентирането в информацията [2], което налага разработването на интелигентни системи за подпомагане на вземането на решения, адаптирани към индивидуалните потребности на потребителя [2], [11]. Макар съвременните методи за анализ на данни и машинно обучение да позволяват откриване

на скрити закономерности, тяхната ефективност зависи от контекста и качеството на използваните данни [12].

В дигиталните библиотеки персонализацията има особено значение поради големия обем дигитализирани документи и ресурси. Интелигентните системи за препоръки анализират съдържанието и потребителското поведение, като осигуряват по-прецизно класифициране и достъп до релевантна информация и повишават ангажираността на потребителите [13], [14].

Подходите за персонализация могат да бъдат статични, основаващи се на предварително зададени настройки, ключови думи или на базата на предварително направени анкети от потребителите [15]. Също така могат да се използват прости статистически методи - генериране на списък с обекти, които представляват интерес за повечето от потребителите или които отразяват определена област на интереси. Другият тип персонализиращи подходи са динамични и използват алгоритми за препоръки, които анализират големи масиви от данни и адаптират съдържанието в реално време. Успешното им приложение във водещи платформи като YouTube, Amazon, Netflix демонстрира потенциала им, а пренасянето им в дигиталните библиотеки представлява логична стъпка към по-ефективна и интерактивна организация и предоставяне на знанието.

2.2.1. Съвременни методи, алгоритми и подходи за персонализация

Съвременните подходи за предоставяне на персонализирано съдържание се базират на адаптивни методи, които в повечето случаи включват системи за препоръки, анализ на потребителското поведение или комбинация от двата подхода. Също така се основават на идеята, че информационната система трябва непрекъснато да се адаптира към променящите се потребности, интереси и поведение на отделния потребител. Тези методи използват интегриран набор от техники и алгоритми за изкуствен интелект и машинно обучение, които работят съвместно, за да осигурят динамично и релевантно представяне на съдържанието.

Основните подходи за генериране на персонализирано съдържание се основават на профилиране и анализ на потребителското поведение, както и на класическите методи за системи за препоръки – филтриране, базирано на съдържание, съвместно филтриране и хибридни решения. Тези подходи често се допълват от методи за класификация и

кълстеризация, които подпомагат откриването на сходства и закономерности и повишават точността на препоръките.

Комбинирането на поведенчески анализ, алгоритми за препоръки и аналитични методи формира адаптивни и обучаващи се системи, способни както да отговарят на текущите потребности на потребителите, така и да предвиждат бъдещи интереси. Въпреки съществуващи ограничения, свързани с качеството и пълнотата на данните, тези подходи представляват ефективна основа за изграждане на персонализирани информационни среди с повишена полезност и ангажираност.

Първият разгледан подход е **анализът на потребителското поведение в уеб среда**, който използва данни от взаимодействията на потребителите със системата, като история на търсенията, навигация и оценки на съдържанието. Чрез обработка и моделиране на тези данни се изграждат потребителски профили, които служат като вход за алгоритмите за препоръки и позволяват адаптирано представяне на релевантно съдържание спрямо индивидуалните интереси и контекст.

В допълнение към подходите, основани на профилиране и анализ на потребителското поведение, широко приложение намират класическите системи за препоръки: филтриране, базирано на съдържание и съвместно филтриране.

Филтрирането, базирано на съдържание, използва сходството между обекти и индивидуалния потребителски профил, като ефективността му зависи от наличието на достатъчно информация за потребителя, както и от качеството на метаданните и семантичните връзки между ресурсите [16]. Основно ограничение на този подход е зависимостта му от вече познатите интереси на потребителя и трудността при откриване на ново съдържание извън установения контекст. Точността на препоръките може да бъде повишена чрез включване на потребителски оценки или индиректни показатели за интерес, като време на взаимодействие със съдържанието, но въпреки това методът остава ограничен при системи с богато и слабо структурирано информационно пространство [17], [18].

Съвместното филтриране преодолява част от тези ограничения, като изгражда препоръки въз основа на сходства между потребители или между елементи, извлечени от модели на потребителско поведение. Чрез анализ на колективните предпочитания се осигурява по-ефективно и мащабируемо персонализиране, особено при наличие на достатъчно данни. Въпреки предизвикателствата, свързани с обработката на големи

обеми информация и избора на подходящи метрики за сходство, този подход се утвърждава като ключов механизъм за предоставяне на персонализирано съдържание в адаптивни информационни системи [17].

В контекста на персонализираните системи и адаптивните препоръчителни подходи, ключова роля играят методите за клъстеризация и класификация дори и само като помощни методи. Разграничават се три основни метода на обучение – **обучение с учител, обучение без учител и обучение с частичен надзор**, които се различават както по наличието и обема на предварителната информация за данните, така и по начина на моделиране на зависимостите между входните и изходните променливи.

Обучението с учител се основава на предварително класифицирани данни и намира широко приложение при задачи за класификация и регресия, включително при обработка на естествен език и тематична категоризация на текстове. Този подход позволява изграждането на модели с висока прогностична точност, но основното му ограничение е необходимостта от големи и надеждно класифицирани обучаващи множества, чието създаване често е ресурсоемко [19]. Сред най-често използваните алгоритми се открояват Наивен Бейсов класификатор, машина на поддържащи вектори и метод на k-най-близките съседи (K-Nearest Neighbors, KNN), отличаващи се с висока точност при прогнозни задачи [20], [21], [22]:

- **Наивният Бейсов класификатор** е вероятностен метод за класификация, основан на теоремата на Бейс и предположението за условна независимост между признаците [23]. Въпреки опростяващото допускане, алгоритъмът показва добра ефективност и висока изчислителна бързина, особено при задачи за класификация на текст и обработка на естествен език [23]. Неговите предимства включват лесна реализация, мащабируемост и стабилна работа при големи и многомерни набори от данни.
- **Машина на поддържащи вектори (МПВ)** е в един от най-популярни подходи за машинно обучение с учител, при който не се използват никакви предварителни знания за проблемната област [24]. МПВ работи много добре с данни с голяма размерност, като избягва „проклятието на размерността” [24]. Той използва само част от обучаващи примери – така наречени поддържащи вектори за представяне на повърхнина на решение [24]. Проблемите, които МПВ решава са обикновено класически задачи за класификация, при които имаме два класа [24].

- Методът на **k-най-близки съседи** (k-NN) е непараметричен и „мързелив“ подход, при който не се строи изричен модел: за нов обект се откриват k най-близки примера в обучаващия набор според избрана метрика, а решението се извежда чрез гласуване (при класификация) или усредняване/претеглено усредняване (при регресия) [25], [26]. Ефективността на k-най-близки съседи зависи пряко от представянето на данните, избора на мярка за близост и стойността на k [27].

Клъстеризацията е основна техника за **обучение без учител**, чиято цел е групиране на данните в сходни клъстери без предварително зададени етикети, като по този начин се откриват вътрешни структури и закономерности в набора от данни [28]. Един от най-широко използваните методи е **k-средните** (k-means), който разделя обектите в предварително определен брой клъстери чрез итеративно минимизиране на вътрешноклъстерното разсейване. Ефективността му силно зависи от избора на броя клъстери. При данни с неясни или припокриващи се граници между групите, естествено продължение на този подход представлява методът на **неясните k-средни** (Fuzzy k-means), който допуска степенувана принадлежност на обектите към повече от един клъстер и осигурява по-гъвкаво и реалистично моделиране на сложни, шумни и динамични данни, характерни за адаптивни и персонализирани системи [29].

Междинен и все по-широко прилаган подход представлява **обучението с частичен надзор**, което комбинира ограничен набор от класифицирани данни с голямо количество неклассифицирана информация [30]. Тази стратегия намалява зависимостта от скъпи процеси по аотиране и осигурява по-добър баланс между точност и приложимост, особено в домейни като класификация на документи и обработка на естествен език.

2.2.2. Представяния на текст и мерки за близост при персонализация

В контекста на персонализацията, особено при прилагането на методи от машинното обучение и изкуствения интелект, ключово значение има не само избраният метод, но и изборът на подходящо представяне на текста и мярка за близост, тъй като обработката и анализът на данните изискват те да бъдат представени в числова форма, което налага прилагането на техники за векторизация и формализиране на текстовата информация.

Утвърдените подходи за векторизация са:

- **Броячното представяне (CountVectorizer)** преобразува текстовите данни в числов вид чрез изграждане на речник от уникални термини и изчисляване на честотата на срещане на всеки термин във всеки документ [31]. Резултатът е разредена матрица „документи–термини“, която може да бъде разширена с n-грамни признаци за улавяне на локални зависимости [31]. Методът е бърз, прозрачен и подходящ като базов модел, но не отчита семантични връзки между термините, поради което често се комбинира с техники за намаляване на размерността.
- **Хеширащото кодиране (HashingVectorizer)** е техника за извличане на признаци, при която токените се съпоставят чрез хешираща функция към индекси в предварително фиксирано пространство, като резултатът е разредена матрица от честоти [32]. За разлика от броячното представяне, методът не поддържа явен речник, което осигурява по-добра мащабируемост и ефективност на паметта и го прави подходящ за големи или потокови данни [32]. Основните му ограничения са невъзможността за обратна интерпретация на признаците и рискът от хеш-сблъсъци, които могат да бъдат намалени, но не и напълно елиминирани чрез избор на достатъчно голямо хеш-пространство [32].
- **Претегленото представяне (TF-IDF, TfidfVectorizer)** е метод за числово представяне на текст, който оценява значимостта на даден термин спрямо конкретен документ и цялата колекция чрез комбиниране на честотата на термина в документа (TF) и обратната честота на документите (IDF) [33]. Мярквата потиска често срещаните в корпуса думи и акцентира върху редките, но характерни за даден документ термини, което я прави ефективна за тематично разграничаване и изчисляване на сходство между документи [33]. В резултат се изгражда разредена матрица „документи–термини“ с TF-IDF тегла, която често превъзхожда простото броячно представяне при задачи за извличане на информация и текстова класификация.
- **Вграждането на думи (embeddings)** представлява усъвършенстван подход за числово представяне на текст, при който се обучават плътни векторни представления, улавящи семантичните и контекстуалните зависимости между думите [34], [35]. За разлика от класическите разредени модели, статичните вграждания присвояват фиксиран вектор на всяка дума, докато контекстуалните модели, базирани на трансформър архитектури (напр. BERT и неговите

производни), генерират представяния, зависещи от конкретната употреба на думата [35], [36]. Макар да изискват значителни изчислителни ресурси и да са по-трудни за интерпретация, методите за вграждане се утвърждават като стандарт в съвременната обработка на естествен език поради високата си семантична изразителност и ефективност при анализ и сравнение на текстове [37].

Популярни метрики за близост, които се ползват, са косинусова близост и коефициент на Жакар, поради тяхната простота и ефективност [38].

ГЛАВА 3. МОДЕЛИ И СОФТУЕРНИ КОМПОНЕНТИ ЗА ПЕРСОНАЛИЗИРАНО ПРЕДСТАВЯНЕ НА СЪДЪРЖАНИЕ В ДИГИТАЛНИ БИБЛИОТЕКИ

В тази глава се представя методологичната основа и архитектурната организация на предложеното решение за представяне на персонализирано съдържание под формата на текстове със сходно съдържание и на персонализирани препоръки в дигитална библиотека. Под персонализация се разбира интегрирането на съдържателни, поведенчески и семантични показатели, посредством които системата адаптира представянето на информационните ресурсите към контекста на ползване и установените предпочитания на конкретния потребител, като едновременно осигурява прозрачни основания за препоръките.

Подходът е **хибриден** и е насочен към **текстови ресурси от тип периодични многотематични издания на български език**. Той интегрира обработката на информационни ресурси, регистри на взаимодействията на потребителите със системата и обогатени метаданни в единна архитектура, при която изчислително тежките стъпки се изпълняват като отделни процеси извън оперативния поток в **асинхронен слой**. Той включва два основни процеса: **семантично представяне на текстовете** с изграждане на **матрица на сходство** между документите и **анализ на потребителските взаимодействия** чрез конструиране на **разредена матрица „потребител–документ“** и **вектор на популярност**. Данните се обработват инкрементално и периодично в пакетен режим, което позволява ефективно обновяване само на новопостъпилата информация и осигурява мащабируемост при големи колекции.

Същинската персонализация и по-точно **интерактивният слой** стъпва върху предварително изчислените оперативни структури, което гарантира ниска латентност, последователност на резултатите и възпроизводимост. Той реализира два вида функционалности: показване на **„подобни документи“** на базата на матрицата на сходство и **генериране на персонализирани препоръки** чрез комбиниране на индивидуални предпочитания и глобална популярност. Моделът цели да преодолее ограничения като „студения старт“, зависимостта от обемни поведенчески данни и високата изчислителна сложност.

3.1. Концептуален модел и архитектурна рамка за персонализирано представяне на съдържание в дигитална библиотека

Предложеното хибридно решение се основава на многослойна архитектура, при която изчислително интензивните операции се изпълняват в етап на предварителна подготовка (асинхронен слой), а интерактивният слой използва предварително изчислени представяния и агрегирани показатели. Тази организация цели да осигури мащабируемост чрез изнасяне на тежките изчисления извън критичния път.

В етапа на предварителна подготовка се ползват **текстовите ресурси и регистрите на потребителските взаимодействия**, допълнени от извлечени **именувани същности**, за да се конструират компактни и инкрементално обновяеми структури, извън непосредственото взаимодействие с потребителя. В този етап текстовият корпус се подлага на унифицирана предварителна обработка и се преобразува в компактни векторни представяния. Върху тях се изгражда матрица на сходство между документите, която обобщава както глобалната тематична близост, така и локални съвпадения, като по избор се обогатява с информация от извлечени именувани същности. Тази матрица служи едновременно като основа за функционалността „подобни документи“ и като ядро на препоръчителния алгоритъм.

Също така се прави анализ на регистрите за потребителски взаимодействия, от които се конструира разрежена матрица „потребител–документ“ и индикатор за глобална популярност. Хибридният алгоритъм комбинира съдържателна близост, индивидуална история и популярност, като осигурява адекватни препоръки дори и при ограничени данни за потребителя.

Интерактивният слой работи изцяло върху тези предварително изчислените структури и реализира два основни режима: „подобни документи“, базиран на смислова близост между текстовете, и „персонализирани препоръки“, които комбинират съдържателна близост, индивидуална история на потребителя и глобална популярност. Така се гарантират ниска латентност, стабилност при нарастване на обема от данни и обяснимост на резултатите.

3.2. Услуга за извличане и структуриране на именувани същности

Ключов елемент от архитектурата е услугата за извличане на именувани същности, която обогатява представянето на документите с допълнителна структурирана семантична информация и се използва при определянето на степента на сходство между

многотематични документи. Нейната основна функция е да **обогаत्या метаданните** на документите чрез *структурирани указатели към лица, организации, географски наименования и други значими обекти, които надграждат стандартните векторни представяния и осигуряват допълнителен семантичен контекст*. По този начин услугата участва в адресирането ключови ограничения, идентифицирани в литературния преглед, като изгражда оценка, която не се базира на единичен компонент, а комбинира няколко допълващи се източника на информация. Това допринася за по-стабилни резултати, смекчава ефекта на „студен старт“, повишава обяснимостта и осигурява мащабируемо семантично обогатяване на метаданните.

Услугата за извличане на именувани същности е реализирана като самостоятелна **асинхронна услуга** в пакетен режим, която обработва документите на партии с минимална предварителна обработка, за да не се наруши разпознаването на същностите. Извлечените именувани същности се съхраняват като обогатени метаданни и подпомагат търсенето, оценката на сходство и генерирането на препоръки. По този начин услугата действа като свързващ компонент между текстовото съдържание, метаданните и персонализиращия алгоритъм.

3.3. Матрица на сходство и метод на многокомпонентна оценка на сходство

Тази подглава представя в детайли основната оперативна структура за моделиране на смислова близост в корпуса - **матрицата на сходство между документите** - и обосновава метода на многокомпонентната оценка. Входната информация за генерирането на матрицата идва от два взаимнодопълващи се източника: (1) **текстовете**, и (2) извлечените от тях **именувани същности**. Върху това се дефинира многокомпонентна оценка, а резултатът се организира като **матрица на сходството**. Така конструираната структура едновременно осигурява обяснимост и оперативна ефективност и служи като основа както за навигация чрез „подобни документи“, така и за формиране на персонализирани препоръки. За да се създаде матрицата на сходство, данните минават през няколко етапа:

- **Подготовка на текстовите данни.** Етапът на подготовка на текстовите данни цели да осигури надеждна основа за изчисляване на сходство чрез редуциране на шума и нормализиране на текста. Процесът включва почистване от нерелевантни символи, уеднаквяване на регистъра, лематизация за намаляване на разредеността и премахване на стоп-думи. По избор се прилага синонимно обогатяване за

ограничаване на лексикалните вариации и подобряване на стабилността на мерките за близост.

- **Векторизация на текстови данни.** Целта на векторизацията е да трансформира подготвените текстове в числови представяния, подходящи за алгоритмите от изкуствения интелект и машинното обучение. Класическите векторизации като „чанта от думи“ и TF-IDF предлагат прозрачно, но повърхностно представяне, тъй като не отчитат семантичната близост между думите. Това ограничение се преодолява чрез вграждания, които представят думите като вектори в многомерно пространство и улавят семантични и контекстуални зависимости, като контекстуалните модели осигуряват по-висока точност при операции по сходство и препоръчване. За векторизация е използван подход, базиран на контекстуални вграждания от трансформър архитектура, поддържащ български език и осигуряващ баланс между качество и изчислителна ефективност. Документите се сегментират в припокриващи се фрагменти с последваща агрегация на ниво документ, което запазва тематичната структура и гарантира стабилни и възпроизводими представяния при инкрементална обработка.
- **Намаляване на размерността при текстови данни.** Документите в дигиталните библиотеки често са дълги, многотематични и със значително семантично припокриване, което затруднява измерването на сходство и увеличава изчислителната сложност. Обичайно решение е прилагането на техники за намаляване на размерността с цел редуциране на броя признаци при запазване на съществена семантична информация, като по този начин се ограничават шумът и излишните корелации и се улеснява интерпретацията [39]. Въпреки предимствата на линейни и нелинейни методи като PCA, NMF и UMAP, тяхната приложимост е ограничена в динамична среда, тъй като проекционните модели зависят от първоначалното разпределение на данните и изискват пълно преизчисляване при тематични промени в корпуса. *Поради това в дисертацията е възприет алтернативен подход, при който вместо допълнителна математическа редукция се използва модел за кодиране, който поначало генерира компактни векторни представяния. Избраният модел MiniLM създава 384-измерни плътни вектори, които значително намаляват изискванията за памет в сравнение с класическите разредени представяния, без загуба на семантична разделителна способност. Този подход позволява линейно и*

инкрементално добавяне на нови документи без необходимост от преизчисляване на индексите, което го прави по-подходящ за реални дигитални библиотеки.

Върху векторизираните представяния на текстовете се дефинира набор от *допълващи се компоненти за оценка на сходството* между документи i и j , целящи да обхванат различни аспекти на смисловата близост:

- **Глобалното сходство** ($S_{\text{усреднено}}(i, j)$) осигурява *обобщена и устойчива мярка за тематично сходство*. Сходството между документите се оценява чрез косинусова мярка, приложена върху техните векторни представяния, като по-високата стойност отразява по-голяма степен на съдържателна близост.
- **Локалното сходство** ($S_{\text{най-добро}}(i, j)$) цели улавяне на силни частични съвпадения на съдържание на ниво фрагмент/сегмент. Подходът е особено релевантен при дигитални периодични издания, където един документ обединява хетерогенни материали (рубрики, статии, новини) с различна тематика. Мярката привилегирова двойки документи, които споделят ясно разграничени тематични блокове, дори когато глобалната им тематика се разминава.
- **Тематичното сходство** ($S_{\text{тематично}}(i, j)$) въвежда *по-груба, но интерпретируема структура чрез групиране в широки тематични области*. При него всички фрагментни вектори се групират чрез алгоритъм за размито клъстеризиране (fuzzy c-means) в C тематични центъра. Всеки фрагмент получава степен на принадлежност към темите.
- Сходството по **именувани същности** ($S_{\text{именувани същности}}(i, j)$) добавя *структуриран семантичен сигнал*, особено информативен при периодични издания. За всеки документ i се формира множество E_i , където E_i е множеството от именуваните същности. Сходството по метаданни се задава чрез коефициента на **Жакар** - този компонент дава допълнителна близост на документи, които споделят съществени общи обекти.

Линейната комбинация от тези компоненти формират балансирана, точна и обяснима оценка на сходството, като всеки допринася различна перспектива към смисловата близост. Всеки компонент се нормализира до съпоставим мащаб, претегля се чрез набор от коефициенти и се сумира. Така се формира компактно разрежено представяне, съгласувано с глобалния индекс на документите. Отделните компонентни

оценки, изчислени върху векторните представяния на текстовете, се обединяват в следната многокомпонентна мярка, отразяваща съдържателната близост:

$$S_{\text{съдържателна}}(i,j) = \alpha * S_{\text{усреднено}}(i,j) + \beta * S_{\text{най-добро}}(i,j) + \gamma * S_{\text{тематично}}(i,j), \alpha + \beta + \gamma = 1$$

Финалната оценка се формира като се добави и компонентът от именувани същности:

$$S_{\text{финална}}(i,j) = (1 - \lambda) * S_{\text{съдържателна}}(i,j) + \lambda S_{\text{именувани същности}}(i,j). \quad 0 \leq \lambda \leq 1$$

На базата на тази многокомпонентна оценка, включваща семантична близост, локални съвпадения, тематични профили и именувани същности, се изчисляват финалните оценки за близост между документите, като получената матрица се съхранява под формата на разреден индекс на **k-най-близки съседи (k-NN)**, което ограничава използваната памет и ускорява достъпа [25]. В този вид тя изпълнява двойна роля: непосредствен вход към модула „подобни текстове“ и опорна основа за последващо генериране на „персонализирани препоръки“.

Хипотезата, която ще бъде валидирана в следваща глава, е, че за текущия корпус от информационни ресурси – многотематични периодични издания – тази многокомпонентна оценка ще може по-добре да улавя сходствата и различията между ресурсите.

3.4. Матрица „потребител-документ“ и имплицитни оценки

Регистрите на взаимодействията между потребителите и системата представлява вторият основен източник на информация в предложената архитектура, наред с текстовите ресурси. Целта на обработката им е да се получи надеждно и компактно представяне на реалното поведение, което може да се използва като имплицитна оценка за интерес и да служи като вход към хибридният модел за генериране на персонализирани препоръки.

Както и при матрицата на сходство, за да се създаде матрицата „потребител-документ“, регистрите за достъп минават през няколко стъпки на обработка:

- **Филтриране и изчистване на регистрите.** Първата стъпка по обработка на регистрите на взаимодействия на потребителите със системата включва изчистване на нерелевантни записи – тъй като се използват само събитията от тип „достъп до ресурс“ се изключват административни операции (създаване, промяна, изтриване на обекти), системни събития и т.н.. След предварителното изчистване

всеки запис се редуцира до минимален набор от полета: идентификатор на потребителя (u), идентификатор на документа (i), времеви печат и тип действие.

- **Агрегиране на имплицитни оценки.** За всеки потребител u и документ i се отчита броят на преглежданията $c_{u,i}$ (заявки от тип „преглед на ресурс“). Поради липса на експлицитни оценки този показател се тълкува като имплицитен индикатор за интерес и се трансформира в имплицитна оценка $w_{u,i}$ чрез монотонно нарастваща, но насищаща се функция. Така единичното преглеждане задава базова тежест; допълнителните преглеждания я повишават с намаляващ прираст. На основата на тези оценки се конструира разредена матрица на взаимодействията.
- **Изграждане на вектор на глобална популярност на документите.** Успоредно с индивидуалните взаимодействия се изчислява вектор на глобална популярност на документите, базиран на броя преглеждания, трансформирани отново чрез монотонно нарастваща, насищаща се функция и нормализиран в интервала $[0,1]$. Този показател се използва като допълнителен фактор в препоръчителния модел, особено при ограничена индивидуална история, като функционира като устойчив глобален сигнал, смекчаващ ефекта на „студения старт“, без да замества семантичната близост и се задава с формулата: $\text{популярност}(i) = \sum_u w_{u,i}$.

3.5. Оперативни структури и механизми за актуализация

Предложеното решение трябва да функционира в динамична среда, в която непрекъснато се добавят, променят или изтриват текстови ресурси и се натрупват нови потребителски взаимодействия със системата. Необходимо е да се гарантират съгласуваността на идентификаторите и на предварително изчислените оперативни структури (матрица на сходство, матрица „потребител-документ“ и вектор на глобална популярност) както и надеждното им съхранение и процедури за периодично/инкрементално обновяване, така че системата да запазва коректност, устойчивост и производителност при нарастващ обем данни и натоварване.

Поради тази причина двата основни потока в системата – съдържателният и поведенческият – се съгласуват чрез обща система от устойчиви идентификатори, която осигурява еднозначно представяне на документите във всички оперативни структури. *Така семантичната близост между текстовете и наблюдаваното потребителско*

поведение могат да се комбинират последователно в рамките на общ препоръчителен модел, без риск от несъответствия между различните представяния.

Системата регистрира всички типове взаимодействия, не само за достъп до ресурс, а и събития като създаване, изтриване и промяна на ресурс, които се ползват за **поддържане на актуалността и консистентността на представянията**. Инкременталното обновяване се реализира чрез обработка само на *новопостъпилите записи*, което позволява добавяне, редактиране или премахване на документи чрез локални актуализации на засегнатите структури, *без необходимост от пълно преизчисляване*.

С нарастването на обема от данни се прилага комбинация от постепенно добавяне и по-рядко *пълно преизчисляване*, когато това е необходимо за възстановяване на максимална последователност. Изчислително тежките операции се изпълняват извън обслужването на потребителските заявки, като интерактивният слой работи единствено с предварително изчислени матрици и индекси. По този начин се осигуряват кратко и предсказуемо време за отговор, възпроизводимост на резултатите и мащабируемост при динамично изменящ се корпус и нарастващ брой потребители.

3.6. Модули за генериране на персонализирано съдържание

Представена са модули за персонализирано препоръчване в дигитални библиотеки, които обединява семантична близост между документите, наблюдавано потребителско поведение и глобална популярност в единна, параметризируема рамка. Модулите целят да осигурява релевантни и обясними препоръки дори и при ограничена индивидуална история, като същевременно поддържа оперативна ефективност и мащабируемост чрез използване на предварително изчислени структури.

При обработка на данните за създаване на оперативните структури *последователността на обработка на данните допуска успоредно изпълнение, но методически е целесъобразно първо да се извлекат именувани същности чрез отделната услуга, за да може последващата матрица на сходство да включи и този допълнителен семантичен принос*. Регистрите за взаимодействия със системата могат паралелно да се обработват със създаването на матрицата на сходство, за да се създаде разредена матрица „потребител–документ“ и вектор на глобална популярност.

Един общ индекс за документите се използва както при матрицата на сходство, така и при матрицата потребител–документ, което осигурява съгласуван преход между навигация чрез „подобни документи“ и персонализирано препоръчване.

Функционалният модул за „подобни документи“ е потребително-инвариантен и се основава изцяло на матрица на сходство между текстовете, която комбинира глобална и локална семантична близост, както и допълнителен принос от именувани същности. Извличането е просто и бързо: системата *локализира активния документ по идентификатор, извлича съответния ред/списък със съседни, изключва самия документ и прилага праг и/или ограничение по k, за да върне най-релевантните резултати*. Тъй като се работи с предварително изчислени стойности, времето за отговор е кратко, а обяснимостта е висока - основанията за предложенията могат да се проследят чрез споделени теми, съвпадащи фрагменти и/или общи именувани същности.

Функционалният модул за генериране на персонализирани препоръки интегрира съдържателната близост с индивидуалната история на потребителя и индикатор за глобална популярност. При потребители с достатъчно натрупани взаимодействия препоръките се формират чрез пренасяне на интересите към семантично близки документи. При **нови или анонимни потребители** се разчита предимно на **индикатора за популярност**, като с натрупване на взаимодействия същият потребител плавно преминава към пълния хибриден режим без промяна на архитектурата. При **слабо активни потребители** *хибридна оценка гарантира стабилност на резултатите: съдържателният показател остава водещ и компенсира оскъдните поведенчески връзки, като извежда смислово близки предложения дори при малък обем история*. При съвсем **нови документи** („студен старт“ за елементи) без натрупани взаимодействия същият *съдържателен компонент предотвратява изолацията им* - те се включват в препоръките на читатели, чиито досегашни текстове са семантично близки до новия документ, като приносът им се регулира с по-ниско тегло до появата на достатъчна поведенческа информация. Оценката на релевантност на персонализираните препоръки се задава със следната формула:

$$\text{оценка}(u, d) = \sum_{i \in \text{история}(u)} w_{u,i} * S_{\text{финална}}(i, d) + \partial(u) * \text{популярност}(d)$$

Хипотезата, която ще бъде проверена в следващата глава, е че тази хибридна оценка генерира по-качествени персонализирани препоръки - особено в гранични сценарии като „студен старт“, оскъдна история и тематични преходи - в сравнение с

решения, основани единствено на поведенчески или единствено на съдържателни показатели.

3.7. Обяснимост и етични принципи при селекция на персонализирано съдържание

В контекста на персонализацията в дигиталните библиотеки все по-голямо значение придобиват прозрачността, обяснимостта и етичното управление на данните, които често остават на заден план спрямо точността на препоръките [40], [45], [41]. Този установен пропуск обосновава необходимостта тези аспекти да бъдат интегрирани още на архитектурно ниво при проектирането на персонализиращи системи.

В предложената архитектура обяснимостта е заложена в самия дизайн на хибридният модел чрез разложима многокомпонентна оценъчна функция, която ясно отделя приноса на семантичната близост, тематичното припокриване, именуваните същности. Това позволява генериране на експлицитни и интуитивно разбираеми обяснения за всяка препоръка, без използване на външни интерпретационни модели [40], [42], [43]. За по-добра възприемаемост числовите оценки се трансформират в лингвистични етикети, калибрирани емпирично и съобразени с изследванията за когнитивната достъпност на ХАИ подходите [44], [45].

Резервна стратегия, използваща глобално популярни документи, предотвратява усилването на пристрастия при оскъдни данни [46], както и чрез функции на насищане, които предотвратяват доминирането на единични интензивни взаимодействия [41]. Паралелно се прилагат принципи за минимизация и защита на данните чрез ограничен обем регистри, псевдонимизация, целево използване на извлечените именувани същности и проследим контрол на достъпа, в съответствие с добрите практики за обясними и отговорни системи [44], [47], [48], [49], [50].

ГЛАВА 4. ЕКСПЕРИМЕНТАЛНО ВНЕДРЯВАНЕ И АНАЛИЗ НА РЕЗУЛТАТНОТО ТЕСТВАНЕ

Главата представя реализацията на модел за персонализирано представяне на съдържание в дигитална библиотека, разработен в съответствие с архитектурата от глава 3. Персонализацията се основава на комбиниран анализ на съдържателната близост между документите, потребителското поведение и допълнителна семантична информация от извлечени именуванни същности. Описани са основните архитектурни компоненти, включително модул за формиране на матрица на сходство, модул за обработка на потребителските взаимодействия и изчисляване на показатели за популярност, услуга за извличане на именуванни същности, както и функционални модули за предоставяне на „подобни документи“ и персонализирани препоръки. В следващите подраздели се разглеждат използваните структури от данни, алгоритми и експерименталната верификация на модела върху синтетични и реални данни.

4.1. Изграждане на технологична среда, тестови данни и протокол за експериментална верификация

Програмната реализация на предложения модел е изградена върху модулна архитектура, оптимизирана за високопроизводителни матрични изчисления и обработка на текст на български език. Технологичният стек е избран с оглед ефективна работа с големи корпуси, възпроизводимост на резултатите и интеграция между асинхронните изчислителни модули и интерактивния слой. Основният език за разработка е Python 3.13+, поради широкото му приложение в машинното обучение и наличието на специализирани библиотеки.

За реализацията на алгоритмичните компоненти се използват sentence-transformers (MiniLM) [51], за генериране на семантични векторни представяния, scikit-learn [52] за предварителна обработка и метрики за сходство, както и scikit-fuzzy [53] за тематично моделиране чрез Fuzzy C-Means. Лингвистичната нормализация на българския текст се осъществява със simplemma [54] и набор от стоп-думи на български език [55].

Изчислителното ядро на системата се базира на NumPy [56] и SciPy [57] за линейна алгебра и разредени структури (CSR), включително при изчисляване на коефициента на Жакар върху множества от именуванни същности, както и на PyTorch [58] за тензорните изчисления на Transformer моделите. Реализацията поддържа хардуерно

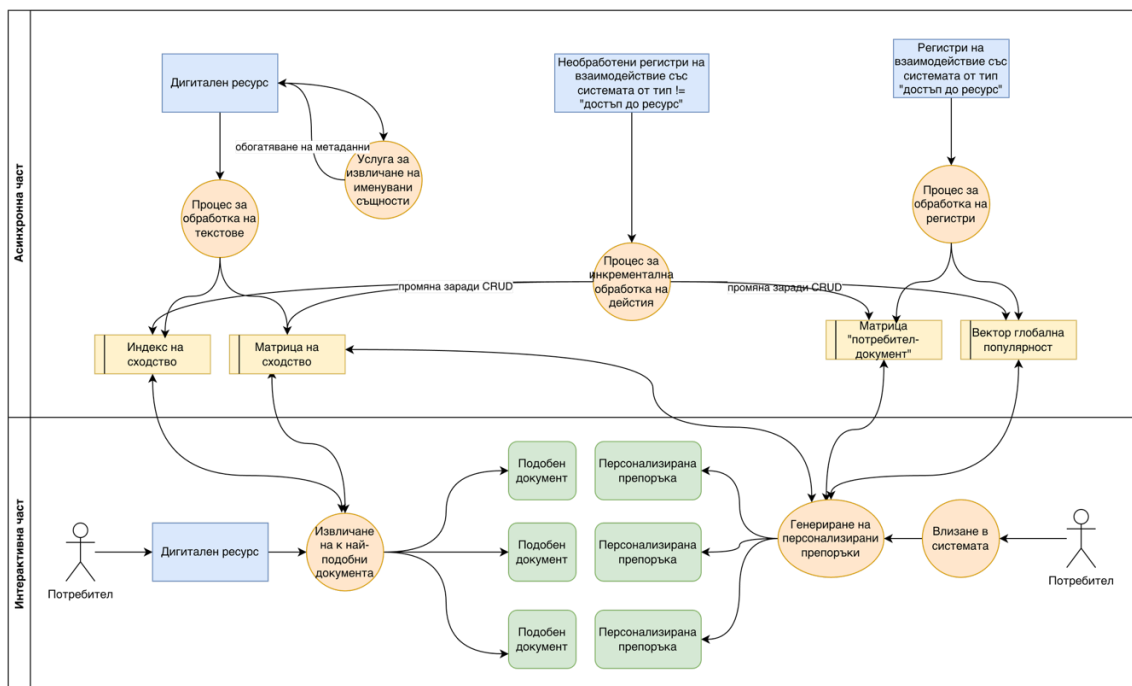
ускорение чрез CUDA и MPS, като семантичните вектори се съхраняват в полупрецизен формат (Float16) за намаляване на паметния отпечатък при запазване на точността.

Емпиричната оценка се базира на корпус от 1000 текстови периодични многотематични ресурси от дигиталната библиотека на Народна библиотека „Иван Вазов“ – Пловдив [59], както и на синтетични набори за валидация на многокомпонентната оценка за сходство и на модула за персонализирани препоръки. Бяха изградени и синтетични потребителски профили, които покриват основни сценарии като „студен старт“, динамична промяна на интересите и ограничена история.

4.2. Архитектура на системата

Архитектурата следва принципа на ясно разделение между изчислително тежките процеси и лека интерактивна част за обслужване на заявки. Тя е структурирана в два слоя – **асинхронен** и **интерактивен слой** (виж фиг. 16).

Асинхронният слой обединява всички изчислително интензивни процеси, чрез които текстовият корпус и регистрите на взаимодействията се преобразуват в предварително изчислени представяния и индекси, използвани от интерактивната част. Това разделяне гарантира ниска латентност, предсказуемо поведение и устойчиво мащабиране при нарастващи обеми от данни и потребителска активност.



Фигура 16. Архитектура на модулите за предоставяне на персонализирано съдържание

В асинхронния слой функционират три независими компонента, синхронизирани чрез общи идентификатори: услуга за извличане на именувани същности, която обогатява документите със структурирана семантична информация; модул за генериране на матрица на сходство, който изгражда разреден индекс на най-близките документи въз основа на семантични и тематични признаци; и модул за обработка на взаимодействия, който конструира разредена матрица „потребител–документ“ и вектор на глобална популярност.

След като асинхронният слой подготви и актуализира всички необходими представяния и индекси, интерактивният слой реализира директното обслужване на потребителските заявки, като работи изцяло върху тези предварително изчислени структури. Интерактивният слой е реализиран чрез два функционални модула. Модулът „подобни документи“ извлича за текущия документ най-близките по съдържание ресурси от матрицата на сходство. Модулът „персонализирани препоръки“ комбинира информация за съдържателната близост между документите, индивидуалната потребителска история и глобалните показатели за популярност, като изключва вече разглежданите ресурси. При липса на достатъчна потребителска история се използват представителни популярни ресурси.

4.3. Услуга за извличане и структуриране на именувани същности

В рамките на архитектурата е реализирана самостоятелна услуга за извличане на именувани същности, чиято цел е да обогатява метаданни на документите със структуриран контекст (напр. лични имена, организации, географски обекти).

Извлечените списъци от именувани същности се използват в две основни направления: като допълнителен семантичен показател при изграждане на матрицата на сходство между документи и като средство за подпомагане на търсенето. По този начин именуваните същности повишават едновременно прецизността и обяснимостта на препоръките, както и навигацията в колекциите. Услугата работи асинхронно в пакетен режим, извън критичния път на потребителските заявки.

Текстовете се обработват с минимални трансформации, тъй като агресивното почистване би влошило качеството на разпознаване на именувани същности. Подходът се базира на предварително обучени трансформър-модели, подбрани за добра покриваемост на български език. Проведена е селекция и експериментално сравнение

върху извадка от документи, като са отчетени както точността, така и времето за обработка.

Алгоритъмът за извличане на именувани същности сегментира дългите документи в застъпващи се фрагменти, от които последователно се извличат същностите и впоследствие се обединяват. Емпиричният анализ показва, че за сегментацията размер от около 200 токена постига оптимален баланс между изчислителна ефективност и достатъчен контекст, докато по-дългите входи водят до спад в производителността. С цел повишаване на устойчивостта и ефективността се прилага двумоделен подход, при който се комбинират резултатите от два модела с най-добро съотношение между точност и бързодействие.

След извличането се извършва постобработка, включваща нормализиране на регистъра, обединяване на резултатите от двата модела чрез максимуми по честота и увереност, както и филтрация с конфигурируеми прагове за брой срещания и увереност. Получените същности се съхраняват като обогатени метаданни, съгласувани с идентификаторите на документите и подготвени за директна интеграция в оценката на сходство и търсенето.

В обобщение, услугата за извличане на именувани същности осигурява устойчиво и мащабируемо обогатяване на документите със структуриран семантичен контекст. Комбинацията от сегментиране, емпирично избран размер на фрагментите, двумоделен анализ и целенасочена постобработка гарантира високо качество на извличане, ниска латентност и добра интеграция с останалите компоненти на системата.

4.4. Матрица на сходство и метод на многокомпонентна оценка. Функционален модул за селектиране на „подобни документи“

С цел повишаване на устойчивостта към лексикални вариации е използван синонимен речник за български език, извлечен от Open Multilingual WordNet [60]. Синонимното обогатяване се прилага единствено в етапа на предварителна обработка при функционалността „подобни документи“, чрез добавяне на ограничен брой представителни синоними преди сегментирането на текстовете. По този начин се намалява чувствителността към повърхностни лексикални различия, без влияние върху интерактивния слой и потребителските обяснения.

Модулът за изчисляване на сходство между текстови документи представлява ядро както на функционалността „подобни документи“, така и на персонализираното

препоръчване. Той функционира в асинхронния слой и е проектиран с фокус върху интерпретируемост, проследимост и изчислителна ефективност, като цели изграждане на надеждно и мащабируемо представяне на смисловата близост. Входът включва нормализиран текстов корпус и извлечени именувани същности за всеки документ, а изходът е разрежена симетрична матрица на сходство, представена като списък от топ-k съседи и съпътстващ речник за съпоставяне на идентификатори.

Процесът по генериране на матрицата започва със зареждане на текстовия корпус и речника на документните идентификатори, последвано от унифицирана подготовка на текста – нормализиране, лематизация, филтриране на стоп-думи и, по избор, синонимно обогатяване. Дългите документи се сегментират в застъпващи се фрагменти, съобразени с максималната допустима дължина на входа за езиковия модел, като застъпването запазва смисловата непрекъснатост и предотвратява загуба на информация по границите.

Всеки фрагмент се кодира в 384-мерен вектор чрез многоезичния модел MiniLM, след което на ниво документ се извеждат агрегирани представяния под формата на усреднени вектори. При активиран тематичен слой се изчисляват и профили на принадлежност, получени чрез размита клъстеризация върху сегментните вектори. Паралелно се зареждат извлечените именувани същности и се конструира структурирано представяне, което служи като допълнителен вход за оценяване на близостта.

Крайната оценка за сходство между два документа се дефинира в интервала $[0, 1]$ и се получава чрез многокомпонентна линейна комбинация на съдържателния компонент, който включва глобално сходство между усреднените вектори, локално сходство чрез максимално съвпадение между сегменти и тематично сходство, базирано на вероятностните профили от размитото групиране. Този семантичен слой се допълва от фактологичен показател, основан на припокриването на именувани същности.

Получената матрица на сходство се материализира в компактни, разреждени структури, оптимизирани за бърз достъп, и се съхранява заедно с речник за двупосочна съпоставимост между документните идентификатори и матричните индекси. Тази матрица се използва за функционалния модул „подобни документи“ в интерактивния слой, което се свежда до директно достъпване, филтриране и подреждане по стойност.

Качеството на матрицата на сходство и тегловните коефициенти (α , β , γ , λ) на отделните компоненти са калибрирани чрез многоетапна експериментална процедура,

комбинираща параметрична оптимизация, аблационен анализ и качествена оценка върху контролирани извадки с нарастващ обем.

За целите на валидацията е конструиран „златен стандарт“, който комбинира лексикално припокриване и припокриване на именувани същности. Валидирането се извършва върху малки, средни и контролни извадки (200, 500 и 1000 документа), като за всяка извадка се изпълняват оптимизация на параметрите, изграждане на матрицата на сходство и независими проверки. Анализът включва структурна верификация на матрицата, измерване на лексикална близост и проверка на семантична свързаност чрез споделени именувани същности, което е особено показателно за тематична близост в периодични издания.

Експерименталните резултати показват, че оптималната стойност за теглото на именуваните същности е $\lambda=0.5$, което дефинира симетричен хибрид между семантичен и фактологичен слой. Това се обосновава със спецификата на периодичния печат, при който информационната стойност е силно концентрирана около конкретни лица, организации и локации. Калибрирането на останалите параметри води до $\alpha = 0.47$, $\beta = 0.07$ и $\gamma = 0.47$, което показва водещата роля на глобалната семантика и тематичното моделиране и второстепенното значение на локалните съвпадения. Така се потиска шумът от повтарящи се рубрики, реклами или стандартни пасажии, характерни за периодичните издания.

Аблационният анализ разкрива последователна динамика: базовият семантичен модел осигурява широк обхват, добавянето на тематичен слой повишава селективността, а включването на именуваните същности възстановява валидни връзки чрез ясен фактологичен контекст. Крайната многокомпонентна оценка постига баланс между прецизност, устойчивост и обяснимост, което я прави подходяща за многотематични дигитални библиотеки. Допълнителна експериментална проверка със синтетичен корпус потвърждава, че пълната многокомпонентна оценка превъзхожда базовите конфигурации по вътрешна кохезия, разграничаване на несвързани теми и улавяне на частични връзки.

4.5. Разредена матрица „потребител-документ“, хибриден алгоритъм и функционален модул за генериране на „персонализирани препоръки“

В тази подглава се валидира поведението на модула за персонализирани препоръки, базиран на елементно-ориентирано съвместно филтриране, в ключови и

гранични сценарии. За изолиране на случайни влияния е използван синтетичен набор от документи, организиран в три тематични клъстера (А, Б и изолиран В), като между А и Б е въведена силна съвместна срещаемост, а клъстер В остава без поведенчески връзки. Допълнително е включен глобално популярен документ, използван като резервен ориентир при липса на история.

Процедурата по верификация обхваща следните сценарии:

1. студен старт за потребител, при който системата коректно преминава към неперсонализирани препоръки по глобална популярност;
2. изчерпан изолиран клъстер, при който при липса на нови релевантни елементи се активира резервна стратегия;
3. смяна на интерес, при която единични взаимодействия извън доминиращия клъстер служат като сигнал за насочване към нова тема;
4. смесен профил с балансирана история, водещ до пропорционално представени препоръки от различни клъстери без „изтичане“ към изолирани теми;
5. студен старт за елемент, при който съдържателният компонент позволява релевантно включване на нов ресурс, макар и с по-ниска оценка;
6. разреждана поведенческа история, при която съдържателният показател компенсира слабите съвместни връзки.

Верифицирането комбинира качествена проверка на логиката на препоръките и количествен анализ на разпределението на резултатите в първите k позиции. Резултатите показват, че модулът генерира предвидими и коректни персонализирани препоръки, запазва полезност при липса или оскъдност на данни и реагира устойчиво при динамична промяна на потребителските интереси.

4.6. Ограничения и валидност на предложената архитектура

Предложената архитектура и хибридният модел за препоръчване са разработени с оглед спецификата на наличния корпус, типа регистри на взаимодействията и използваните методи за обработка на текст. За коректна интерпретация на резултатите е необходимо ясно разграничаване на основните ограничения и обхвата на валидност на модела. От гледна точка на данните, моделът е най-надежден при многотематични периодични издания на български език, като пренасянето му към други езици или тематични домейни изисква повторна калибрация и допълнителна емпирична оценка.

Качеството на текстовото съдържание представлява съществен фактор – OCR грешки, липсващи сегменти и т.н. внасят шум, който пряко влияе върху векторните представяния, сходствата и извлечените именувани същности и не може да бъде напълно елиминиран.

Съдържателният модел е ограничен от зависимостта си от предварително обучени езикови модели, които не са оптимизирани за всички домейни и могат да пропускат редки, остарели или специфични термини. Използването на именувани същности като допълнителен индикатор зависи от надеждността на модула за разпознаване, като грешки или пропуски могат да повлияят върху оценките за сходство, макар че филтрирането по честота и увереност смекчава този ефект. Многокомпонентната оценка изисква избор на прагове и тегловни коефициенти, които се определят емпирично и влияят върху крайната структура на сходствата.

Поведенческите данни са с имплицитен характер: всеки преглед се интерпретира като положителен сигнал, без гаранция за реално прочитане или удовлетворение. Това въвежда несигурност, която се компенсира чрез комбиниране със съдържателни показатели и популярност, но не може да бъде напълно отстранена. Разпределението на взаимодействията е силно неравномерно, като част от документите остават с оскъдна история, а анонимните потребители се третират агрегирано.

Техническите ограничения произтичат от изчислителната тежест на векторизацията, изграждането на матрицата на сходство и тематичните модели, което налага пакетна обработка, ограничаване на съхраняваните сходства и внимателен избор на честота за обновяване. Използването на големи езикови модели за извличане на именувани същности увеличава хардуерните изисквания и може да изисква специализирана инфраструктура. Също така препоръките не отразяват всяко ново взаимодействие в реално време поради пакетния режим на обработка - компромис между актуалност и изчислителна приложимост.

По отношение на валидността, вътрешната валидност е подкрепена от ясно дефинирани и повторяеми процеси за подготовка на данните и изграждане на оперативните структури, което осигурява възпроизводимост на резултатите. Външната валидност е ограничена до среди със сходна структура на съдържанието и регистрите на взаимодействията и при прилагане в различен контекст изисква адаптиране на параметрите и повторна оценка. Разделът очертава границите, в които предложената архитектура е коректна и полезна, и служи като основа за реалистична интерпретация на резултатите и планиране на бъдещи подобрения.

ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

Научни приноси:

1. Разработени са концептуален модел и архитектурна рамка за персонализирано представяне на съдържание в дигитална библиотека. Те включват асинхронен слой за изграждане на оперативни структури (матрица на сходство, матрица „потребител-документ“, вектор на глобална популярност на документите, структури от именувани същности) и интерактивен слой с ниска латентност. Формализирани са оперативните структури и релациите между тях, което осигурява възпроизводимост, проследимост и съвместимост между модулите.
2. Предложен е метод за многокомпонентна оценка за сходство между многотематични документи, дефинирана като линейна комбинация от показатели за глобална семантична близост, локални съвпадения на фрагменти, тематични профили и именувани същности. Оценките за сходство между документите са запазени в матрица на сходство, която едновременно обслужва функционалността „подобни документи“ и подпомага генерирането на персонализирани препоръки.
3. Разработен е хибриден алгоритъм за генериране и предоставяне на релевантно съдържание спрямо нуждите на потребители в дигитална библиотека. Той е елементно ориентиран и агрегира съдържателната близост между кандидат-документите и елементите от индивидуалната история на потребителя върху предварително изчислена матрица на сходство. За осигуряване на стабилност е въведен резервна стратегия, базирана на вектор на глобална популярност на документите, който се прилага в гранични случаи като „студен старт“ и оскъдна история. По този начин се гарантира полезността на препоръките дори при ограничен брой наблюдавани взаимодействия.

Научно-приложни приноси:

1. Реализирана е услуга за извличане и структуриране на именувани същности от текстове на български език като допълнителен информационен показател, която обогатява описателните данни на документите, подобрява оценката за близост и подпомага търсенето по структурирани полета.
2. Реализиран е функционален модул за селектиране на „подобни документи“, който генерира персонализирано съдържание върху параметризируема

многокомпонентна оценка за близост между документи. Модулът осигурява извличане на най-близки съседи с ниска латентност, инкрементални обновявания без преизчисляване на целите оперативни структури и пълна съвместимост с интерактивния слой на системата.

3. Реализиран е елементно ориентиран хибриден алгоритъм, който стъпва на разредена матрица „потребител-документ“. Той е ядрото на функционален модул за персонализирани препоръки. Използва се за генериране на препоръки за „подобни документи“, близки до текущата история на взаимодействие на потребителя. Кандидат-документите се формират от съседи по близост, изключват се вече прегледани ресурси и се подреждат чрез съчетаване на съдържателен принос и имплицитни тежести. При необходимост, се ползва глобална популярност на документите в случаи на „студен старт“ или при изчерпани персонализирани препоръки.
4. Дефинирани и внедрени са механизми за периодична актуализация и инкрементално допълване на оперативните структури и наблюдаваните взаимодействия, осигуряващи мащабируемост и устойчивост на системата при нарастващи обеми от данни и интензивност на потребителската активност.
5. Извършена е експериментална валидация на функционален модул „подобни документи“. Проведени са систематично претърсване на параметрите и аблационен анализ на многокомпонентната оценка (семантика, локални съвпадения, тематични профили, именуван същности), с цел да се калибрират теглата и да се оцени индивидуалният принос на всеки компонент към крайния резултат.
6. Извършена е експериментална валидация на алгоритъма и функционалния модул за генериране на персонализирани препоръки. Изпълнени са контролирани сценарии, обхващащи ключови гранични случаи (студен старт за потребител и за елемент, разредени данни, смяна и смесване на интереси, изчерпване на препоръки). Емпирично е доказано, че хибридното агрегиране на съдържателна близост, потребителска история и популярност води до предвидимо и устойчиво поведение на препоръчващия модул, включително в граничните сценарии.

АПРОБАЦИЯ

Част от представените резултати в настоящото дисертационно изследване са постигнати и апробирани при участието на автора в дейностите по научни проекти и програми:

- Национална научна програма „Млади учени и постдокторанти – 2“, одобрена с РМС № 206 / 07.04.2022 г.
- Национална интердисциплинарна изследователска Е-инфраструктура за ресурси и технологии за българското езиково и културно наследство, интегрирана в рамките на европейските инфраструктури CLARIN и DARIAH (КЛАДА-БГ), Министерство на образованието и науката, програма „Националната пътна карта за научна инфраструктура“, 2018-2027, изпълнители: ИИКТ-БАН (координатор), ИМИ-БАН (партньор), ДО1-167/28.07.2022, ДО1-324/01.12.2023, ДО1-97/26.06.2025.

Списък с авторски публикации по темата на дисертацията

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

Списък с докладвани резултати

1. Изнесен научен доклад на международна конференция Computer Vision Conference 2026, 21-22 Май 2026, Амстердам, Нидерландия, Тема: A Multi-Component Similarity Measure for Personalized Content Discovery in Periodical Digital Library Collections, Автори: **E. Mitreva**, D. Paneva-Marinova, V. Georgiev, A. Nikolova, Дата: 22.05.2026
2. Изнесен научен доклад на международната конференция Computer Systems and Technologies 2024 (CompSysTech '24), 14–15 Юни 2024, Русе, България, Тема: Classification of Short Noisy Text, Автори: **E. Mitreva**, V. Georgiev, A. Nikolova, Дата: 15.06.2024
3. Изнесен научен доклад на международната UNESCO конференция Digital Presentation and Preservation of Cultural and Scientific Heritage, 07–10 Септември 2023, Бургас, България, Тема: Personalization Approaches for Cultural Heritage Study, Автори: **E. Mitreva**, A. Nikolova, V. Georgiev, A. Gigova, Дата: 07.09.2023
4. Изнесен научен доклад на Годишна отчетна сесия на секция „Математическа лингвистика“ на ИМИ-БАН на 22.12.2023, София, България, Тема на доклада: Подходи за персонализация при изучаване на културното наследство, Автор: **Е. Митрева**
5. Изнесен научен доклад на Годишна отчетна сесия на секция „Математическа лингвистика“ на ИМИ-БАН на 18.12.2024, София, България, Тема на доклада: Класификация на кратки зашумени текстове, Автор: **Е. Митрева**

6. Изнесен научен доклад на Годишна отчетна сесия на секция „Математическа лингвистика“ на ИМИ-БАН на 10.12.2025, София, България, Тема на доклада: Модели и методи за предоставяне на персонализирано съдържание в дигитални библиотеки, Автор: **Е. Митрева**

Списък с цитирания

Общо открити цитирания – 4 (без автоцитирания).

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)

Цитирана в:

- 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

Цитирана в:

- 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),

Цитирана в:

- 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

БИБЛИОГРАФИЯ

- [1] Z. Liu и B. Shao, „A systematic review of library services platforms research and research agenda,” *Library & Information Science Research*, том 46, № 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*, том 25, № 3, pp. 39-58, 2025. <https://doi.org/10.48550/arXiv.2506.15055>.
- [3] J. C. Shikali и 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 и L. Zlatkov, „CultIS: Web-based Platform for Intelligent Cultural Content Management,” *Digital Presentation and Preservation of Cultural and Scientific Heritage*, том 14, pp. 19-36, 2024. <https://doi.org/10.55630/dipp.2024.14.1>.
- [5] M. Goynov, D. Luchev, D. Paneva-Marinova, R. Pavlov и K. Rangochev, „Towards Providing Analytical Services in the Web-Based Platform for Intelligent Cultural Content Management CultIS,” *TEM Journal*, том 14, № 4, pp. 2946-2952, 2025. doi: 10.18421/TEM144-05.
- [6] C. G. S. и M. Mulimani, „The Impact of Artificial Intelligence on Library and Information Science (LIS) Services,” *Social Science Research Network*, том 14, № 5, pp. 50-56, 2024. doi: <http://dx.doi.org/10.2139/ssrn.4856459>.
- [7] D. Paneva-Marinova, M. Goynov и R. Pavlov, „Enhanced and personalized learning experience in digital libraries,” в *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 и 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*, том 9, № 1, pp. 4830-4842, 2025. doi: 10.47772/IJRISS.2025.9010372.
- [9] V. A. Kumar и M. Chidambaram, „Personalization and User Behavior Analysis in Digital Libraries: A Systematic Review,” *Academic Research Journal of Science and Technology (ARJST)*, том 2, № 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*, том 1, № 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 и A. Hotho, „Semi-supervised Learning: An In-depth Parameter Analysis,“ в *Lecture Notes in Computer Science*, том 12873, 2021. https://doi.org/10.1007/978-3-030-87626-5_5.
- [20] R. S. Nurhalizah, R. Ardianto и P. Purwono, „Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review,“ *Jurnal Ilmu Komputer dan Informatika*, том 4, № 1, pp. 61 - 72, 2024. doi: 10.54082/jiki.168.
- [21] E. Najjar и A. M. Breesam, „Supervised Machine Learning a Brief Survey of Approaches,“ *Al-Iraqia Journal of Scientific Engineering Research*, том 2, № 4, p. 71–82, 2023. doi: 10.58564/IJSER.2.4.2023.121.
- [22] X. Zhang, F. Guo, C. Tao , P. Lei, G. Beliakov и J. Wu, „A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research,“ *Journal of Theoretical and Applied Electronic Commerce Research*, том 18, № 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 (icWCSN '23)*, 2023. doi: <https://doi.org/10.1145/3585967.3585984>.
- [25] Y. Sun и Q. Liu, „Collaborative filtering recommendation based on K-nearest neighbor and non-negative matrix factorization algorithm,“ *The Journal of Supercomputing*, том 81, № 79, 2024. doi: <https://doi.org/10.1007/s11227-024-06537-4>.

- [26] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal и A. Khraisat, „Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications,” *Journal of Big Data*, том 11, 2024. doi: <https://doi.org/10.1186/s40537-024-00973-y>.
- [27] D. R. Anamisa, A. Jauhari и F. A. Mufarroha, „K-Nearest Neighbors Method for Recommendation System in Bangkalanâ s Tourism,” *ComTech: Computer, Mathematics and Engineering Applications*, том 14, № 1, pp. 33-44, 2023. doi: 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*, том 30, pp. 110-123, 2024. doi: 10.1016/j.ecosta.2021.11.008.
- [30] S. Li, G. Yuan, M. Yang, Y. Shen, C. Li, R. Xu и X. Zhao, „Improving Semi-Supervised Text Classification with Dual Meta-Learning,” *ACM Transactions on Information Systems*, том 42, № 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 и J. Nandigam, „A Survey of Text Representation and Embedding Techniques in NLP,” *IEEE Access*, том 11, pp. 36120-36146, 2023. doi: 10.1109/ACCESS.2023.3266377.
- [35] Z. Nie, Z. Feng, M. Li, C. Zhang, R. Zhang, D. Long и R. Zhang, „When text embedding meets large language model: a comprehensive survey,” *arXiv preprint arXiv:2412.09165*, 2024. doi: 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.
- [37] K. Das, Kamlish и F. Abid, „Advancements in Word Embeddings: A Comprehensive Survey and Analysis,” *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, том 61, № 3, 2024. doi: 10.53560/PPASA(61-3)842.
- [38] A. Levy, B. R. Shalom и M. Chalamish, „A guide to similarity measures and their data science applications,” *Journal of Big Data*, том 12, № 1, 2025. doi: 10.1186/s40537-025-01227-1.
- [39] A. K. Rastogi, S. Taterh и B. S. Kumar, „Dimensionality reduction algorithms in machine learning: a theoretical and experimental comparison,” в *International Conference on Recent Advances in Science and Engineering*, 2023. doi: <https://doi.org/10.3390/engproc2023059082>.
- [40] Y. Zhang и X. Chen, „Explainable Recommendation: A Survey and New Perspectives,” *Foundations and Trends® in Information Retrieval*, том 14, № 1, pp. 1-101, 2020. doi: <https://doi.org/10.1561/15000000066>.
- [41] S. Milano, M. Taddeo и L. Floridi, „Recommender Systems and Their Ethical Challenges,” *AI & Society*, том 35, p. 957–967, 2020. doi: <https://doi.org/10.1007/s00146-020-00950-y>.
- [42] Tintarev, N. и Masthoff, J., „Designing and Evaluating Explanations for Recommender Systems,” в *Recommender Systems Handbook*, 2021, pp. 479–510. doi: https://doi.org/10.1007/978-0-387-85820-3_15.
- [43] F. Doshi-Velez и 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*, том 267, p. 1–38, 2019. doi: <https://doi.org/10.1016/j.artint.2018.07.007>.
- [45] M. T. Ribeiro, S. Singh и 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*, том 159, № 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 и F. Herrera, „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,“ *Information Fusion*, том 99, 2023. doi: <https://doi.org/10.1016/j.inffus.2023.101805>.
- [50] J. Z. Forde и 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,“ [Онлайн]. Available: <https://huggingface.co/sentence-transformers>.
- [52] „Machine Learning in Python,“ [Online]. Available: <https://scikit-learn.org/stable/>.
- [53] „scikit-fuzzy,“ [Онлайн]. Available: <https://pypi.org/project/scikit-fuzzy/>.
- [54] A. Barbaresi, „Simplemma: a simple multilingual lemmatizer for Python,“ [Онлайн]. Available: <https://github.com/adbar/simplemma>.
- [55] „Stopwords Bulgarian (BG),“ [Онлайн]. Available: <https://github.com/stopwords-iso/stopwords-bg>.
- [56] „NumPy,“ [Онлайн]. Available: <https://pypi.org/project/numpy/>.
- [57] „SciPy,“ [Онлайн]. Available: <https://scipy.org/>.
- [58] „PyTorch,“ [Онлайн]. Available: <https://pytorch.org/>.
- [59] „Дигитална библиотека Народна библиотека "Иван Вазов" Пловдив,“ [Онлайн]. Available: <https://digital.libplovdiv.com/bg>.
- [60] B. Group. [Онлайн]. Available: <https://github.com/omwn/omw-data/tree/main/wns/bul>.