



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

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

Дисертационен труд

на

Емануела Димитрова Митрева

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

СЪДЪРЖАНИЕ

Съдържание	2
Благодарности	4
Увод	5
Глава 1. Обща постановка на задачата	8
1.1. Обект, предмет, цел и задачи на изследването	8
Глава 2. Теоретични основи и анализ на съвременни подходи за персонализация в дигитални библиотеки	10
2.1. Дигитални библиотеки - същност и развитие	11
2.1.1. Концептуални и функционални аспекти на дигиталните библиотеки	11
2.1.2. Разграничение между дигитални и традиционни библиотеки	13
2.1.3. Ролята на потребителя и необходимостта от персонализация	15
2.2. Персонализация в дигитални библиотеки	16
2.2.1. Съвременни методи, алгоритми и подходи за персонализация	18
2.2.2. Сравнение, предизвикателства и ползи от различните подходи	31
2.2.3. Представяния на текст и мерки за близост при персонализация	35
2.2.4. Преглед на съвременни изследвания и използвани методи за персонализация в дигитални библиотеки	41
Глава 3. Модели и софтуерни компоненти за персонализирано представяне на съдържание в дигитални библиотеки	59
3.1. Въведение и обхват	59
3.2. Концептуален модел и архитектурна рамка за персонализирано представяне на съдържание в дигитална библиотека	62
3.2.1. Концептуален модел	62
3.2.2. Архитектурна рамка	67
3.3. Услуга за извличане и структуриране на именувани същности	70
3.4. Матрица на сходство и метод на многокомпонентна оценка на сходство	71
3.4.1. Подготовка на текстовите данни	71
3.4.2. Векторизация на текстови данни	72
3.4.3. Намаляване на размерността при текстови данни	73
3.4.4. Матрица на сходство и вход към „подобни текстове“	76
3.5. Матрица „потребител-документ“ и имплицитни оценки	79
3.5.1. Филтриране и нормализиране на събитията	79
3.5.2. Агрегиране на имплицитни оценки	79
3.5.3. Вектор на глобална популярност на документите	80
3.6. Оперативни структури и механизми за актуализация	80
3.7. Модули за генериране на персонализирано съдържание	81
3.7.1. Функционален модул за „подобни документи“ и метод на многокомпонентната оценка	82
3.7.2. Функционален модул за персонализирани препоръки и хибриден алгоритъм	83
3.8. Обяснимост и етични принципи при селекция на персонализирано съдържание	85

3.8.1. Архитектурна обяснимост и интерпретация на резултатите.....	86
3.8.2. Минимизиране на алгоритмичните пристрастия и защита на данните	87
3.9. Обобщение	87
Глава 4. Експериментално внедряване и анализ на резултатното тестване.....	90
4.1. Изграждане на технологична среда, тестови данни, протокол за експериментална верификация ...	90
4.2. Архитектура на системата	93
4.2.1. Асинхронен слой за подготовка и актуализация на оперативните структури.....	94
4.2.2. Интерактивен слой за бързо обслужване на заявки	96
4.3. Услуга за извличане и структуриране на именувани същности	96
4.4. Матрица на сходство и метод на многокомпонентна оценка. Функционален модул за селектиране на „подобни документи“	100
4.4.1. Синонимно обогатяване на текстовото представяне (OMW-Bulgarian Wordnet).....	100
4.4.2. Реализация на функционалния модул.....	101
4.4.3. Експериментална валидация на функционален модул „подобни документи“.....	108
4.5. Разредена матрица „потребител-документ“, хибриден алгоритъм и функционален модул за генериране на „персонализирани препоръки“	119
4.5.1. Реализация на модула.....	119
4.5.2. Експериментална валидация и тестови сценарии на модула за генериране на персонализирани препоръки	125
4.6. Обяснимост и етични механизми в реализацията	132
4.6.1. Обяснимост на алгоритмите	132
4.6.2. Принципи за защита и минимизация на данните	133
4.7. Ограничения и валидност на предложената архитектура	134
4.7.1. Ограничения, свързани с данните.....	134
4.7.2. Ограничения на съдържателния модел.....	134
4.7.3. Ограничения на поведенческите данни.....	135
4.7.4. Технически и изчислителни ограничения.....	135
4.7.5. Валидност на резултатите	136
4.8. Обобщение	137
Глава 5. Заключение – резюме на получените резултати	139
5.1. Обобщение на резултатите	139
5.2. Насоки за бъдещо развитие	140
Приноси на дисертационния труд	142
Списък на авторските публикации по темата на дисертацията.....	144
Списък на цитирания	145
Списък на докладвани резултати.....	147
Списък на фигурите	148
Списък на таблиците.....	149
Речник на използваните термини и съкращения.....	150
Библиография	155
Декларация за оригиналност на резултатите.....	167

БЛАГОДАРНОСТИ

На първо място бих искала да изразя своята най-искрена благодарност към научния ми ръководител проф. д-р Десислава Панева-Маринова за подкрепата, насоките, търпението и доверието през целия процес на работа по настоящата дисертация. Благодаря за професионалното менторство, градивната критика и многобройните ценни дискусии, които допринесоха не само за реализирането на този труд, но и за моето научно развитие. Особено съм благодарна за постоянната вяра в мен и в завършването на този труд - дори в моментите, когато самата аз трудно намирах увереност, че ще стигна до края.

Специални благодарности изразявам и към доц. д-р Максим Гойнов, който оказва съществена помощ чрез предоставянето на необходимите данни за проведените тестове, както и чрез няколко изключително ползотворни разговора помежду ни. Именно те поставиха основата и дадоха първоначалния тласък за идеята, върху която впоследствие се разви настоящото дисертационно изследване.

Искрено благодаря и на моето семейство и приятели за търпението, разбирането и подкрепата през целия период на работа по дисертацията. И най-вече сърдечни благодарности отправям към братовчед ми и жена му, които не само ме мотивираха да започна докторантура, но и неизменно ме подкрепяха през годините с идеи, съвети, мотивация и искрена вяра в мен и работата ми.

УВОД

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, образованието и обществената комуникация. В този контекст дигиталните библиотеки заемат особено място като среди, в които се обединяват дългосрочното съхранение, надеждното описание и организираното 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 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 за документите и потребителите и различни подходи, с цел постигане на по-устойчиво, обяснимо и приложимо в реална среда препоръчване.

Настоящият дисертационен труд е посветен на изследване и разработване на модел за персонализирано представяне на съдържание в дигитални библиотеки. Подходът стъпва върху методи на изкуствения интелект и машинното обучение за анализ на текстове, извличане на именувани същности (имена, локации и т.н.) и формиране на препоръки. Основната задача е създаването на архитектурна рамка и софтуерни

компоненти, които да бъдат технически изпълними, съобразени с особеностите на библиотечната среда и подлежащи на емпирична проверка.

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, свързани с представения дисертационен 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 препоръчителните системи и извършва аналитично изследване на научните постижения и резултати от актуалните изследвания и научните 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 разгледаните 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 1. ОБЩА ПОСТАНОВКА НА ЗАДАЧАТА

1.1. 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 начините, 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.

Дисертационния 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 на информационните ресурси, 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. В рамките на изследването се акцентира върху:

- Използване на големи езикови модели за разпознаване и извличане на именувани същности за обогатяване на метаданните на текстовите ресурси.
- Разработване на модели за препоръчване и филтриране на съдържание въз основа на сходства между текстове и потребителски интереси и поведение.
- Интегрирането на тези методи в хибридна препоръчваща система, която подпомага адаптирането на библиотечното съдържание към индивидуалните потребители.

Изследването се основава на хипотезата, че прилагането на подходящи методи на изкуствения интелект и машинното обучение за адаптиране на съдържанието 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 дигитални 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.

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 знанието 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.

За да се отговори на тези предизвикателства, все по-голямо значение придобиват интелигентните методи за персонализация, които целят да адаптират съдържанието според индивидуалните потребности, интереси и поведение на потребителя. Персонализацията се разглежда като ключов подход за подобряване на достъпа до знание, чрез който информационната среда става по-гъвкава, адаптивна и ориентирана към конкретния потребител.

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. Чрез този преглед и анализ се цели очертаването на тенденции, проблеми и предизвикателства, свързани с внедряването на персонализирани услуги.

2.1. Дигитални библиотеки - същност и развитие

2.1.1. Концептуални и функционални аспекти на дигиталните библиотеки

Развитието на дигиталните технологии през последните десетилетия доведе до фундаментална трансформация в начина, по който знанието се създава, съхранява, организира и разпространява. Тази трансформация не се ограничава единствено до дигитализацията на съществуващи информационни ресурси, а обхваща изграждането на нови модели за управление на знанието, основани на мрежови технологии, автоматизация и интелигентни системи. В този контекст появата на дигиталните библиотеки представлява естествен етап от еволюцията на информационната среда. Те се възприемат не просто като дигитализирани колекции, а като интегрирани системи за управление на знания, които съчетават информационни ресурси, технологична инфраструктура и разнообразни услуги в динамична и адаптивна среда [1], [2].

Съвременните дефиниции [3], [4], [5] подчертават, че дигиталната библиотека е многопластова структура, която обединява технологични, организационни и социални измерения. Тя функционира едновременно като информационна система, комуникационна платформа и социална екосистема, в която взаимодействат различни групи потребители - изследователи, преподаватели, студенти и широката общественост. Дигиталните библиотеки се разглеждат като среди, в които „интелигентните технологии и инструментите с изкуствен интелект осигуряват персонализиран достъп и подобро потребителско преживяване“ [6], [7], [8]. В този смисъл дигиталната библиотека може да бъде дефинирана като структурирана информационна система, която съхранява, индексира, обработва и предоставя дигитални обекти чрез мрежови технологии, гарантирайки тяхната устойчивост, интероперативност и дългосрочна достъпност [2], [9].

Функционалната архитектура на дигиталните библиотеки се основава на няколко взаимосвързани направления, които осигуряват тяхната ефективност и устойчивост.

- **Достъп и откритост.** Дигиталните библиотеки елиминират пространствените и времевите ограничения, характерни за традиционните институции. Те предоставят непрекъснат достъп до ресурси независимо от географското местоположение на потребителя. Както се отбелязва в [10], [11], [12], дигиталните платформи играят ключова роля за поддържане на образователния процес и научните изследвания в глобален мащаб, особено

в условия на кризи или ограничена физическа мобилност. Освен това те подкрепят принципите на отворения достъп и свободното разпространение на научни резултати.

- **Съхранение и дългосрочна устойчивост.** Съхранението на дигитални обекти изисква прилагането на стандартизирани протоколи, формати и политики за дългосрочно архивиране. Авторите в [1], [13] посочват, че съвременните дигитални библиотеки използват облачни технологии, разпределени хранилища и механизми за резервиране на данни, за да гарантират тяхната устойчивост във времето. Дългосрочната дигитална консервация се превръща в стратегически приоритет, свързан със запазването на културната и научната памет.
- **Индексиране и управление на метаданни.** Ефективното управление на метаданни е ключова предпоставка за откриваемост, интероперативност и интеграция между различни информационни системи. Стандартизираните схеми за метаданни позволяват свързаност между колекции и улесняват автоматизираното извличане на информация [11]. В този контекст семантичните технологии и свързаните отворени данни допринасят за по-добра структурираност и смислова интеграция на ресурсите.
- **Търсене и интелигентно извличане на информация.** Процесите на търсене и класификация се развиват значително благодарение на изкуствения интелект. Както се отбелязва в [14], алгоритмите за машинно обучение позволяват анализ на съдържанието и потребителското поведение, което подобрява релевантността на резултатите и улеснява навигацията. Интелигентните системи за препоръки подпомагат откриването на нови ресурси и създават по-интерактивна и адаптивна среда.
- **Потребителски профили и персонализация.** Персонализацията се утвърждава като една от най-значимите тенденции в развитието на дигиталните библиотеки. В прегледите на [7], [8], [15] се подчертава, че все по-често се използват механизми за персонализация, базирани на анализ на поведенчески данни, интереси и изследователски профили. Прилагането на техники на изкуствен интелект и машинно обучение

позволява динамично адаптиране на интерфейса, препоръките и услугите към индивидуалните нужди на потребителя. Това води до по-висока степен на ангажираност, ефективност и удовлетвореност от използването на системата [15].

Основната цел на дигиталните библиотеки е да осигурят широк, устойчив и равнопоставен достъп до знание, като подкрепят научните, образователните и културните процеси в глобален мащаб [2], [10]. Те се стремят не само към дигитализация на съдържанието, но и към изграждане на интелигентна информационна екосистема, която интегрира данни, услуги и потребители в динамична среда на взаимодействие [2], [5], [16]. Според авторите на [17] дигиталните библиотеки играят ключова роля в подкрепа на отворената наука чрез осигуряване на интегрирани механизми за споделяне, достъп и повторна употреба на научни данни. Авторите на [18] ги разглеждат като стратегическа инфраструктура за съхранение и разпространение на научна информация, особено в контекста на развиващите се държави, където дигиталните платформи преодоляват географските и икономическите бариери пред знанието. В допълнение, изследванията в [14], [19] подчертават, че чрез внедряване на изкуствен интелект и адаптивни технологии дигиталните библиотеки постигат по-висока степен на персонализация и достъпност, улесняват навигацията и повишават ефективността на изследователската работа.

По този начин значението на дигиталните библиотеки надхвърля традиционната функция на съхранение и предоставяне на информация. Те се превръщат в активен посредник в процесите на създаване, откриване, анализ и споделяне на знание. Като интелигентна и устойчива инфраструктура те подпомагат развитието на иновации, академични практики и съхраняването на културната памет, утвърждавайки се като ключов елемент от съвременната информационна екосистема.

2.1.2. Разграничение между дигитални и традиционни библиотеки

Технологичните и концептуалните разлики между традиционните и дигиталните библиотеки са разгледани в Таблица 1 [2], [20], [21].

Таблица 1. Сравнителен анализ между традиционни и дигитални библиотеки

<i>Аспект на сравнение</i>	<i>Традиционна библиотека</i>	<i>Дигитална библиотека</i>
<i>Тип на ресурс</i>	Съдържа физически носители - книги, периодика, ръкописи и архивни материали.	Включва дигитални обекти - текстови, аудио- и визуални ресурси, бази от данни и мултимедийни формати.
<i>Начин на достъп</i>	Изисква физическо присъствие и достъп в определено работно време.	Осигурява дистанционен и непрекъснат (24/7) достъп чрез онлайн платформи и интерфейси.
<i>Обхват и аудитория</i>	Обслужва локална общност или институционални потребители.	Поддържа глобална достъпност и възможност за обслужване на неограничен брой потребители.
<i>Съхранение и поддръжка</i>	Използва физическо пространство и поддръжка на материални носители.	Прилага сървърни, облачни и репликационни системи за дългосрочно цифрово съхранение.
<i>Организация и индексирание</i>	Каталози и класификационни системи.	Използва автоматизирани методи и стандартизирани схеми за метаданни.
<i>Механизми за търсене</i>	Търсене по основни библиографски признаци (автор, заглавие, ключови думи).	Поддържа семантично и пълнотекстово търсене, филтриране и препоръчителни алгоритми с елементи на изкуствен интелект.
<i>Персонализация и адаптивност</i>	Ограничена - базирана на индивидуално обслужване от библиотекар.	Адаптивна - използва анализ на поведение и автоматизирани механизми за персонализиране на съдържанието.

<i>Аспект на сравнение</i>	<i>Традиционна библиотека</i>	<i>Дигитална библиотека</i>
<i>Взаимодействие с потребителя</i>	Пасивно - потребителят заявява нуждите си и получава резултат.	Активно и интерактивно - системата реагира динамично на потребителското поведение и контекст.
<i>Роля на библиотекаря</i>	Основна оперативна роля в организацията и предоставянето на ресурси.	Еволюира към консултативна и технологична функция - управление на данни и дигитални услуги.
<i>Основни предизвикателства</i>	Физическо остаряване на материалите, ограничено пространство, разходи по поддръжка.	Технологична зависимост, киберсигурност, стандартизация и устойчивост на дигиталното съдържание.

От направеното сравнение между традиционните и дигиталните библиотеки ясно се откроява тенденцията към трансформация от статичен подход за съхранение на ресурси към динамична, технологично управлявана платформа за достъп и взаимодействие с потребителите. Дигиталните библиотеки разширяват обхвата на знанието отвъд физическите граници, осигурявайки глобален, непрекъснат и персонализиран достъп до ресурси [10]. За разлика от традиционните библиотеки, които се фокусират върху съхранението и опазването на физически носители, дигиталните системи поставят акцент върху интеграцията, адаптивността и анализа на потребителското поведение [20]. В резултат, дигиталната библиотека се явява не просто средство за достъп до информация, а активна среда за създаване, споделяне и развитие на знание, което е в основата на съвременните научни и образователни процеси [2], [14].

2.1.3. Ролята на потребителя и необходимостта от персонализация

В съвременните дигитални библиотеки потребителят заема централна и активна роля. Дигиталната среда променя начина, по който потребителите взаимодействат с информационните ресурси, като създава условия за персонален достъп, динамично търсене. Според [15], 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 [22]. В този контекст персонализираните системи в дигиталните библиотеки функционират като интелигентни посредници, които обединяват алгоритми за машинно обучение, анализ на потребителските профили и препоръчващи модели, за да оптимизират достъпа до информацията.

Потребителят вече е не просто статичен консуматор на информация и ресурси, а активен участник в изграждането на знание. Както отбелязват [23], [24], въвеждането на персонализирани услуги подпомага не само ангажираността, но и развитието на критично мислене, като насърчава интерактивното търсене и сътрудничеството между потребители и системи. Персонализацията е особено значима при големи обеми данни, когато информационното претоварване може да затрудни достъпа до релевантни ресурси както е в случая с дигиталните библиотеки.

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 бележи прехода към общество, в което информацията е основен стратегически ресурс. В този контекст способността на организациите и отделните потребители да извличат смисъл и знание от огромни масиви данни се превръща в ключово конкурентно предимство [25]. Увеличаващото се количество, разнообразие и сложност на данните обаче затруднява процеса на ориентиране, анализ и извличане на релевантна информация [2]. Това налага разработването на интелигентни системи за подпомагане на вземането на решения, които да филтрират и представят информацията по начин, адаптиран към индивидуалните потребности на потребителя

[2], [21]. В този контекст персонализацията се явява ключов подход за намаляване на информационната сложност и за повишаване на ефективността при взаимодействието между потребителя и дигиталните ресурси.

Съвременните технологии за извличане на данни, анализ и машинно обучение предоставят възможност за откриване на скрити зависимости и закономерности, но тяхното ефективно прилагане изисква задълбочено разбиране на контекста и доверие в качеството на използваните данни [26]. Често математическите и статистическите допускания, върху които се основават аналитичните модели, са трудно обясними за неспециалисти, което усложнява вземането на рационални решения, основани на данни [26].

С нарастването на обема от данни и ускорената дигитализация на съдържанието се увеличава необходимостта от персонализиран достъп до информация. Дигиталните ресурси често съществуват в различни формати, без ясна категоризация и установени взаимовръзки, което затруднява ефективното им използване. В този контекст персонализацията изпълнява ролята на интелигентен посредник между потребителя и съдържанието, подпомагайки откриването на релевантна информация.

Персонализираните системи прилагат методи на изкуствения интелект и машинното обучение за анализ на съдържателните характеристики на ресурсите и поведенческите данни на потребителите, като изграждат адаптивни профили, актуализирани динамично в зависимост от взаимодействието с информационната система. Това позволява предоставянето на индивидуализирано съдържание, съобразено с интересите и предпочитанията на всеки потребител.

В дигиталните библиотеки персонализацията има ключово значение поради мащаба и разнообразието на дигитализираните документи и научни ресурси. Използването на интелигентни препоръчващи системи подпомага по-прецизното класифициране и предлагане на релевантни материали, подобрявайки достъпа до информация и ангажираността на потребителите [27], [28].

Подходите за персонализация могат условно да се разглеждат като статични и динамични. При статичната персонализация съдържанието се адаптира въз основа на предварително зададени потребителски настройки, докато динамичната се реализира чрез алгоритми за препоръки, които анализират големи масиви от данни, за да откриват закономерности и предвиждат бъдещи интереси. Последните позволяват по-дълбоко

разбиране на контекста и поведението на потребителя, като осигуряват динамично адаптиране на съдържанието в реално време.

В практиката на водещи платформи като Netflix, Amazon и YouTube подобни модели вече са доказали своята ефективност при ангажиране на потребителите чрез индивидуализирани препоръки. Пренасянето на тези принципи в контекста на дигитални библиотеки представлява естествена еволюция в начина, по който знанието се организира и предоставя. Персонализираните системи не само оптимизират достъпа до информация, но и подобряват качеството на взаимодействие между човека и дигиталното знание, като подпомагат изграждането на по-интерактивна и ангажираща информационна среда.

В следващия раздел се обобщават различните подходи за препоръчване на персонализирано съдържание и се посочват разликите, предимствата и недостатъците на тези подходи.

2.2.1. Съвременни методи, алгоритми и подходи за персонализация

Предоставянето на препоръки за релевантно съдържание на потребители може да се извърши по различни начини - статично чрез директно запитване на потребителите за техните предпочитания и/или добавяне на ключови думи в профилите им, които съответстват на техните интереси [29], а също и като се използват прости статистически методи - генериране на списък с обекти, които представляват интерес за повечето от потребителите или които отразяват определена област на интереси. По-прогресивният начин за предоставяне на персонализирано съдържание е чрез използване на адаптивни методи - повечето, от които включват система за препоръки или анализ на действията на потребителите, а дори и комбинация от двете.

Статични методи

Първата група методи, които ще бъдат разгледани, може да бъде условно определена като статични, тъй като при тях информацията за потребителските предпочитания се предоставя директно от потребителя или се събира чрез първоначални проучвания и впоследствие се използва без динамично адаптиране. В ранни изследвания [26], [30] са разгледани възможностите за дефиниране на потребителски настройки, които могат да бъдат актуализирани ръчно от потребителите и предоставят информация за техните интереси. Тези настройки могат да се интерпретират като форма на статична

персонализация, тъй като, въпреки че подлежат на промяна, те не се адаптират автоматично въз основа на потребителското поведение.

Възможно подобрене на този подход е системата да събира информация за обектите, достъпвани от потребителите, и на тази основа да прилага статистически методи за извеждане на списъци с най-често използваните ресурси. Допълнително усъвършенстване, което вече се доближава до адаптивен подход, може да се постигне чрез идентифициране на семантични връзки между ресурсите [31], позволяващи групиране на популярни обекти по тематични области и съпоставянето им с интересите, съхранявани в потребителските профили.

Друг вариант на статичен метод се основава на първоначални проучвания, анкети или тестове [32], чрез които се събират данни за индивидуалните предпочитания на потребителите.

Въпреки че статичните методи не осигуряват динамична адаптация и имат ограничена ефективност в дългосрочен план, те са особено приложими в началните етапи на използване на системата, когато липсват достатъчно поведенчески данни. В този контекст те допринасят за преодоляване на т.нар. проблем на студения старт (cold start), характерен за адаптивните препоръчващи системи. Освен това, както е посочено в [28], потребителите могат да предоставят информация за предпочитания от тях стил на обработка на информация, върху който да се базират препоръките. Тези подходи могат ефективно да се комбинират с адаптивни методи след натрупване на достатъчен обем данни, тъй като повечето адаптивни алгоритми демонстрират висока ефективност при наличие на богата информационна база, но показват ограничени резултати при оскъдни данни.

Адаптивни методи

Адаптивните методи за персонализация се характеризират с динамично моделиране на потребителските предпочитания въз основа на наблюдаваното поведение и контекст на взаимодействие. Те прилагат съвкупност от техники и алгоритми от областта на изкуствения интелект и машинното обучение за автоматично актуализиране на потребителските модели и осигуряване на релевантно представяне на съдържанието в реално време.

Основните подходи за генериране на персонализирано съдържание включват профилиране и анализ на потребителското поведение в уеб среда, както и класически

методи за изграждане на препоръчващи системи - филтриране, базирано на съдържание, съвместно филтриране, базирано на потребители, и съвместно филтриране, базирано на елементи, както и техни хибридни комбинации, насочени към повишаване на точността и адаптивността на препоръките. Класификационните и клъстеризационните методи се прилагат както като самостоятелни подходи, така и като допълващи механизми, които повишават ефективността на водещите алгоритми.

Профилирането и анализът на потребителското поведение представляват фундаментален компонент на персонализацията и се основават на събиране и обработка на данни за взаимодействието между потребителя и системата, включително история на търсенията, кликания, време на престой и оценки на съдържанието. Извлечените поведенчески модели служат като основа за изграждане на механизми за препоръчване, които чрез анализ на сходства между потребители или между обекти подпомагат предоставянето на релевантна информация [33]. Класификационните и клъстеризационните методи допринасят за повишаване на точността на препоръките чрез откриване на закономерности и формиране на групи със сходни характеристики. Съвместното действие на тези подходи позволява моделиране на потребителските предпочитания, отчитане на промените в поведението и прогнозиране на бъдещи информационни потребности.

В следващите подраздели са разгледани основните направления на адаптивните методи, които служат като теоретична и методологична основа за изграждането на интелигентни персонализирани системи за достъп до информация.

Профили и анализ на потребителското поведение в уеб среда (Web Usage Mining)

Персонализацията, базирана на потребителски профили, може да бъде прилагана както в рамките на статични, така и на адаптивни методи, в зависимост от използваните техники. При статичния подход, разгледан в предходния раздел, профилните настройки се дефинират ръчно от потребителя и служат като основа за промяна в представянето, подредбата и нивото на детайлност на съдържанието [26].

В настоящия раздел фокусът е върху подходи, които не разчитат на предварително зададени настройки, а на анализ на данни, събрани от историята на взаимодействията между потребителите и системата. В уеб базираните среди тези данни обикновено се съхраняват в регистри за достъп и действия и, освен за административни

и одитни цели, могат да бъдат използвани за идентифициране на сходства между потребителите и предоставяне на персонализирано съдържание.

Един от основните динамични подходи в този контекст е анализът на потребителското поведение в уеб среда, при който профилите се групират въз основа на сходни поведенчески характеристики с цел предлагане на съдържание, релевантно за съответната група потребители. Този подход е разгледан в редица изследвания [26], [31] и има за цел моделиране и прогнозиране на потребителските интереси чрез анализ на регистрите за взаимодействие [34].

Процесът включва няколко основни етапа (виж фиг. 1): събиране на данни от регистрите за достъп и действия, предварителното им почистване и редуциране, както и извличане на полезни закономерности и модели [35], [36]. Качеството на етапа на предварителна обработка е от критично значение, тъй като значителна част от данните често произтича от невалидни заявки или автоматизирани обхождания [37], [38], което може да доведе до извеждане на нерелевантни правила.



Фигура 1. Етапи на подхода анализ на потребителското поведение в уеб среда

Прилагането на този подход е свързано както с предизвикателства, породени от големия обем данни, така и с проблема на студения старт, характерен за потребители с ограничена или липсваща история на взаимодействия. В такива случаи обикновено се комбинират адаптивни и статични методи за персонализация.

Въпреки присъщите ограничения, анализът на потребителското поведение в уеб среда представлява ценен спомагателен подход, който допринася за повишаване на точността и контекстуалната чувствителност на персонализираните системи.

Алгоритми за филтриране и препоръка – филтриране, базирано на съдържание и съвместно филтриране, базирано на потребители или елементи

Системите за препоръки са широко използвани в електронната търговия за предлагане на продукти въз основа на потребителско поведение и исторически данни [39]. Аналогични подходи могат да бъдат приложени и в системи за управление и достъп

до съдържание в дигитални библиотеки чрез използване на филтриране, базирано на съдържание, и съвместно филтриране, базирано на елементи или потребители.

Филтрирането, базирано на съдържание представлява подход за генериране на персонализирани препоръки, при който се анализира сходството между характеристиките на наличните обекти и потребителски профил, изграден въз основа на предходните взаимодействия и предпочитания на конкретния потребител [40]. За разлика от методите, които препоръчват ресурси единствено въз основа на сходството между самите обекти, при този подход препоръките се формират чрез съпоставяне на характеристиките на обектите с историята на взаимодействията на потребителя. Ефективността на метода зависи както от наличието на достатъчно информация за предпочитанията на потребителя, така и от качеството и степента на структурираност на метаданните и семантичните характеристики на препоръчваните обекти [31]. При ограничени данни за потребителските интереси или липса на явна обратна връзка под формата на оценки съществува риск от препоръчване на ресурси, които са сходни по формални характеристики, но не съответстват на действителните потребности и предпочитания на потребителя [41], [42]. Този недостатък може частично да бъде преодолян чрез използване на имплицитна обратна връзка, като време на взаимодействие с даден ресурс, честота на достъп или други поведенчески индикатори за интерес.

Съвместното филтриране представлява подход за генериране на препоръки въз основа на сходствата в предпочитанията между множество потребители, а не на индивидуален потребителски профил за разлика от филтрирането, базирано на съдържание [40], [41]. Основната хипотеза е, че потребители, проявили сходно поведение в миналото, вероятно ще демонстрират сходни интереси и в бъдеще. В литературата се разграничават два основни подхода – съвместно филтриране, базирано на потребители, и съвместно филтриране, базирано на елементи [42].

Съвместното филтрирането, базирано на потребители, идентифицира потребители със сходни модели на поведение чрез анализ на оценки, предпочитания или история на взаимодействията. Препоръките се формират въз основа на обекти, които са получили положителна оценка или са били използвани от потребители с висока степен на сходство [40], [42]. Ефективността на подхода зависи от наличието на достатъчно данни за потребителските взаимодействия и от възможността за надеждно определяне на сходства между потребителите. Сред основните ограничения се открояват проблемите, свързани с разредеността на данните, динамичния характер на потребителските

предпочитания и нарастващата изчислителна сложност при увеличаване на броя на потребителите [41], [42].

Съвместното филтриране, базирано на елементи, определя сходства между обектите въз основа на моделите на тяхното използване или оценяване от потребителите [42]. Препоръките се генерират чрез идентифициране на обекти, които са сходни с ресурси, към които потребителят вече е проявил интерес [42]. В сравнение с подхода, базиран на потребители, този метод осигурява по-висока стабилност на матриците на сходство, тъй като взаимовръзките между обектите обикновено се изменят по-бавно от потребителските предпочитания - възможно е дадени потребители да си изменят интересите [42]. Това позволява по-ефективно мащабиране и намалява необходимостта от често преизчисляване на сходствата [42]. Ограниченията на подхода са свързани с необходимостта от достатъчно исторически данни за взаимодействията с обектите, както и с понижената ефективност при нови или рядко използвани елементи, за които липсва достатъчна информация за изграждане на надеждни зависимости [42].

И двата подхода на съвместно филтриране са зависими от наличието на достатъчно исторически данни за взаимодействията между потребители и обекти. При ниска плътност на матрицата потребител-обект възниква проблемът с разредеността на данните, който затруднява надеждното изчисляване на сходствата и води до понижаване на качеството на препоръките. Допълнителни предизвикателства са свързани с обработката на големи обеми поведенчески данни и с необходимостта от ефективни механизми за периодично актуализиране на моделите за препоръчване [41], [42].

Въпреки тези ограничения, съвместното филтриране се счита за един от най-ефективните подходи за персонализация [41], особено при наличие на достатъчно богата база от потребителски данни. Както е демонстрирано в системи като ShareTEC [29], практическата му реализация изисква отделно съхранение на поведенческите данни и асинхронна обработка на изчислително интензивните алгоритми, с цел осигуряване на навременни препоръки в реална среда.

Алгоритми за класификация и клъстеризация

В условията на нарастващи обеми от текстови данни ръчната обработка и извличането на знания са ресурсоемки и податливи на субективни фактори. Поради това в съвременните информационни системи се прилагат методи от изкуствения интелект и

машинното обучение за автоматизиране на анализа, повишаване на надеждността на резултатите и ограничаване на информационното претоварване [43].

В машинното обучение се разграничават три основни парадигми - обучение с учител, обучение без учител и обучение с частичен надзор, които се различават по степента на налична предварителна информация и по начина на моделиране на зависимостите в данните.

Обучението с учител използва предварително етикетирани данни и е широко прилагано в задачите по обработка на естествен език, в частност при тематична класификация [44]. То е ефективно при задачи за класификация и регресия, като позволява изграждането на модели с висока прогностична точност. Често използвани алгоритми са k-най-близки съседни, наивен Бейсов класификатор, класификационни дървета и машини на опорните вектори [45], [46], [47]. Основно ограничение на подхода е необходимостта от големи и надеждно етикетирани данни, чието създаване е скъпо и времеемко [48].

Обучението без учител не изисква етикетирани данни и се прилага за откриване на латентни структури и зависимости. Типичен пример е клъстеризацията, при която обектите се групират на базата на вътрешно сходство [49]. Широко използвани техники са методът k-средни и невронните мрежи, които подпомагат проучвателния анализ на данни [45], [50]. Въпреки по-голямата си гъвкавост, този подход обикновено постига по-ниска точност и по-ограничена обяснимост в сравнение с методите с учител.

Обучението с частичен надзор представлява хибриден подход, при който малък набор от етикетирани данни се комбинира с по-голям обем неетикетирани. Това позволява повишаване на точността и намаляване на зависимостта от ръчно етикетирани данни [51]. Подходът е особено приложим в области с висока цена на ръчно етикетирани данни, като медицинската диагностика, класификацията на документи и обработката на естествен език. [52].

Следващата секция представя и анализира класически методи за класификация и клъстеризация на текстове.

Наивен Бейсов класификатор (Naive Bayes)

Наивният Бейсов класификатор представлява приближение на оптималния Бейсов класификатор. Нека H е множеството от хипотези, а D - множеството от обучаващи данни. Оптималната Бейсова класификация на нов пример се определя чрез

агрегиране на предсказанията на всички хипотези, претеглени с техните апостериорни вероятности [53]. За краен набор от класове V апостериорната вероятност примерът да принадлежи към клас $v_j \in V$ се дефинира като:

$$P(v_i|D) = \sum_{h_j \in H} P(v_j|h_j)P(h_j|D)$$

Оптималното Бейсово решение се определя като класът, за който тази вероятност е максимална:

$$v_j \equiv \arg \max_{v_j \in V} \sum_{h_j \in H} P(v_j|h_j)P(h_j|D)$$

Нека примерът x бъде описан чрез вектор от атрибутни стойности $x = \langle a_1, \dots, a_n \rangle$, а целевата функция $f(x)$ приема дискретни стойности от крайното множество V . Задачата за класификация се формулира като намиране на максимално апостериорната оценка:

$$V_{MAP} = \arg \max_{v_j \in V} P(v_j|a_1, \dots, a_n)$$

Съгласно формулата на Бейс това води до:

$$V_{MAP} = \arg \max_{v_j \in V} P(a_1, \dots, a_n|v_j)P(v_j)$$

Вероятността $P(v_j)$ се оценява чрез относителната честота на класа v_j в обучаващото множество. Изчисляването на съвместната вероятност $P(a_1, \dots, a_n|v_j)$ е изчислително сложно, поради което се въвежда предположението за условна независимост на атрибутите при фиксиран клас [53], [54]:

$$P(a_1, \dots, a_n|v_j) = \prod_{i=1}^n P(a_i|v_j)$$

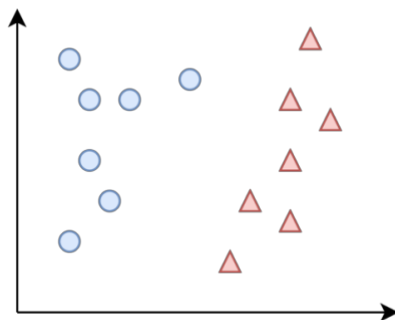
При това предположение правилото за класификация на наивния Бейсов класификатор се задава като::

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i=1}^n P(a_i|v_j)$$

Машина на опорните вектори - МПВ (SVM)

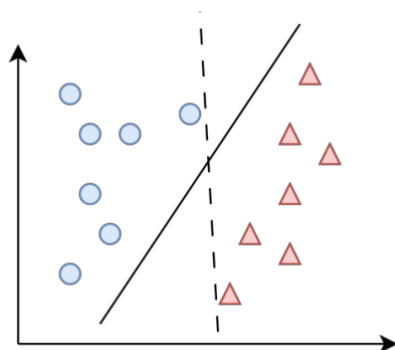
Машината на опорните вектори (Support Vector Machine, SVM) е в един от най-популярни подходи за машинно обучение с учител, при който не се използват никакви

предварителни знания за проблемната област [55]. Методът е особено ефективен при работа с данни с висока размерност, като в значителна степен ограничава ефекта на т.нар. „проклятие на размерността“ [55]. МВП използва подмножество от обучаващите примери - опорните вектори - за дефиниране на повърхността за вземане на решение [55]. Най-често методът се прилага при задачи, при които имаме два класа [56].



Фигура 2. Данни за класификация от Машина на опорните вектори

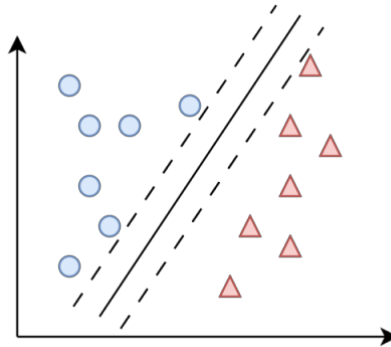
Както е илюстрирано на фигура 2, данните принадлежат към два класа, които могат да бъдат линейно разделени. Поради съществуването на множество възможни разделящи прави, алгоритъмът на SVM цели да определи оптималната разделяща хиперплоскост, наричана повърхност за вземане на решение [56]. Всяка такава повърхност съответства на линеен класификатор [57]. Въпреки че различни класификатори могат да дават идентични резултати за дадено обучаващо множество, тяхното поведение може да се различава при класификация на нови входни данни [56].



Фигура 3. Два класификатора от Машина на опорните вектори

Както е показано на фигура 3, МВП предпочита класификатора с по-голям разделящ интервал, тъй като той осигурява по-добра обобщаваща способност [56]. При

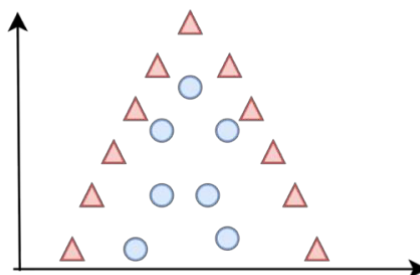
фиксирана ориентация на повърхността за вземане на решение и при отсъствие на грешки в класификацията, оптималната позиция на тази повърхност е тази, която максимизира минималното разстояние до най-близките точки от двата класа [56]. Изместването ѝ извън тези гранични позиции води до неправилно разделяне на данните.



Фигура 4. Хиперравнина в Машина на опорните вектори

Областта между двете успоредни пунктирани линии на фигура 4 дефинира допустимата граница на класификатора, а централната линия между тях представлява оптималната повърхност за вземане на решение при запазване на текущата ориентация [55], [56], [58]. Разстоянието между граничните линии определя интервала на класификация, който следва да бъде максимизиран. Точките, разположени върху тези гранични линии, се наричат опорните вектори и определят позицията на оптималната хиперплоскост [56].

При по-сложни разпределения на данните, каквито са показани на фигура 5, линейното разделяне на класовете може да се окаже невъзможно. В такива случаи не съществува повърхност за вземане на решение, която да осигури пълно и коректно разделяне на примерите, независимо от позицията ѝ [56].



Фигура 5. По-сложни данни за разделяне от Машина на опорните вектори

Метод на k-най-близки съседи (k-NN)

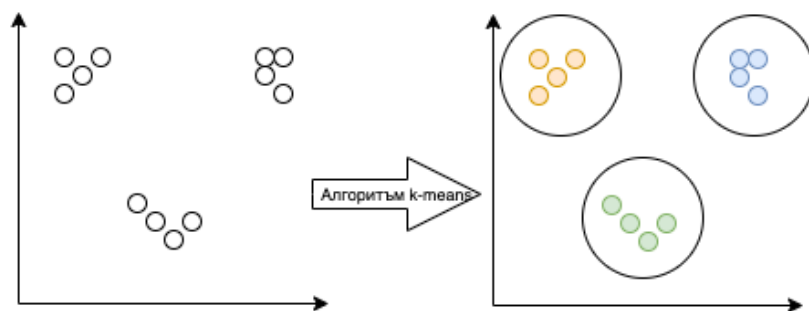
Методът k-най-близки съседи (k-NN) е непараметричен и „мързелив“ подход, при който не се строи изричен модел: за нов обект се откриват k най-близки примера в обучаващия набор според избрана метрика, а решението се извежда чрез гласуване (при класификация) или усредняване/претеглено усредняване (при регресия) [59], [60].

Ефективността на k-най-близки съседи зависи пряко от представянето на данните, избора на мярка за близост и стойността на k [61]. При по-големи k се осигурява по-гладко поведение, но с риск да „размие“ локалната структура [61]. В практиката метриката се съобразява с домейна - Евклидово разстояние или Манхатън при стандартизирани числови вектори, косинусова близост при разредени текстови представяния - а претеглянето на приноса на съседите с функция на разстоянието често повишава устойчивостта в области с неравномерна плътност [62]. Основно предизвикателство е изчислителната цена: наивното търсене изисква сравнение с всички налични примери и става трудно при големи набори и/или висока размерност; затова се използват индекси и приблизително търсене на съседи (ANN), които съществено ускоряват заявките при приемлива загуба на точност [62]. В контекста на обяснимостта k-най-близки съседи е привлекателен, тъй като естествено позволява аргументиране на решението чрез посочване на конкретните k близки примера и техните тежести; при небалансирани класовете са налични варианти с претегляне и адаптивен избор на съседи, които смекчават пристрастията към доминиращия клас [63].

Метод на k-средните (k-means) и метод на размитите k-средни (Fuzzy k-means)

Клъстеризацията е основна техника в машинното обучение за обучение без учител, насочена към организиране на данните в хомогенни групи въз основа на мярка за сходство [64], [65]. Целта е постигане на висока вътрешноклъстерна сходност и максимална междуклъстерна разнородност, което позволява извличане на латентна структура и закономерности в данните без предварително зададени етикети [65].

Един от най-широко използваните методи за клъстеризация е алгоритъмът на k-средните (k-means) [65], [66], поради своята концептуална простота и изчислителна ефективност [67]. Методът разделя набор от n обекта на k неприпокриващи се клъстера ($k \leq n$), като броят на клъстерите се задава предварително [64]. Обектите се разпределят така, че сходството вътре в клъстерите да бъде максимално, а различието между клъстерите - минимално (виж фиг. 6).



Фигура 6. Алгоритъм на к-средни (k-means)

Алгоритъмът функционира итеративно чрез инициализация на k клъстерни центроида, разпределяне на обектите към най-близкия центроид според избрана метрика за разстояние и последващо преизчисляване на центроидите като средни стойности на съответните клъстери [68]. Процесът продължава до достигане на критерий за сходимост, като стабилизиране на клъстерите или минимизиране на средноквадратичната грешка [64], [68], [69].

Нека $x = \{x_1, x_2, \dots, x_n\}$ е множеството от данни, а S_1, S_2, \dots, S_k са клъстерите, а $\mu_1, \mu_2, \dots, \mu_k$ са средните стойности на данните в клъстера S_i [69], [70]. Оптимизацията при метода на k -средните може да бъде формулирана като задача за минимизиране на сумата от квадратичните отклонения:

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

Ефективността на метода в значителна степен зависи от избора на броя клъстери k , което представлява едно от основните му ограничения [65], [71].

При данни с неясно изразени граници между клъстерите или при наличие на припокриване между групите класическият подход може да бъде разширен чрез метода на размитите k -средни (Fuzzy k-means) [72], [73]. За разлика от твърдата кластеризация, този метод допуска степенувана принадлежност на обектите към клъстерите чрез стойности в интервала $[0,1]$, което позволява по-гъвкаво моделиране на неопределеност и преходни състояния [72], [74].

Съществуват разширения на метода, включващи изчисляване на принадлежности на базата на матрици от разстояния [73], както и многоизгледни модели с ентропийна регуларизация [75], които подобряват устойчивостта и качеството на резултатите. Тези

подходи намират приложение при анализ на сложни и шумни данни, характерни за социални мрежи, препоръчващи системи, потребителско поведение и медицинска диагностика.

В този контекст методите за размитата клъстеризация представляват подходящ инструмент за откриване на скрити структури и моделиране на припокриващи се групи, което ги прави особено релевантни за адаптивни и персонализирани информационни системи.

Обучение с частичен надзор (Semi-supervised learning)

Обучението с частичен надзор заема междинно място между обучението с учител и обучението без учител, като комбинира ограничен набор от етикетирани данни с голям обем неетикетирани данни [76], [77]. Подходът позволява повишаване на ефективността на класификационните модели при значително намалена зависимост от ръчно етикетирани обучаващи множества, чието създаване често е скъпо и времеемко [78], [79]. Допълнително предимство е подобрената способност за обобщаване на моделите в сценарии с ограничени етикетирани данни [80], [81], [82].

Една от широко използваните техники в рамките на обучението с частичен надзор е итеративното самообучение, при което моделът първоначално се обучава върху наличните етикетирани данни, след което генерира псевдоетикети за част от неетикетираните примери и ги включва в последващи етапи на обучение. Основното предположение на този подход е, че обекти със сходни характеристики е вероятно да споделят и сходни етикети, което позволява извеждане на допълнителна обучаваща информация чрез анализ на близостта между данните [83], [84].

Ефективността на този метод зависи от изпълнението на редица предпоставки. На първо място, първоначалният набор от етикетирани данни следва да бъде балансиран и представителен за всички класове, тъй като отклоненията в разпределението могат да доведат до натрупване на грешки и влошаване на точността [80], [85], [86]. Съществено значение има и подборът на псевдоетикетите, включвани в обучението, тъй като използването на всички автоматично генерирани етикети може да доведе до пренапасване и загуба на обобщаваща способност [82], [87].

Поради това на практика се прилагат стратегии за избор на псевдоетикети въз основа на праг на увереност, като в обучението се включват само примери с достатъчно висока вероятност за коректност. Използването на статичен праг обаче показва

ограничена ефективност при малки етикетирани множества [87], [88]. По-устойчив подход представлява динамичната адаптация на прага на достоверност, при която първоначално се използват само високонадеждни псевдоетикети, а впоследствие прагът се понижава контролирано с цел по-пълно използване на наличните данни и процесът се повтаря [89]. Повтарящият се процес на актуализация позволява на модела да извлича потенциално ценна информация, скрита в немаркираните примери, което води до по-добри резултати при вземането на решения и по-добра обобщеност на модела и избягване на пренапасване.

Допълнително подобрене може да бъде постигнато чрез претегляне на псевдоетикетите според степента им на увереност, като по-надеждните примери оказват по-силен принос при обучението на модела [90].

В следващия раздел се анализират предизвикателствата и предимствата на различните методи за персонализация, като се аргументира необходимостта от хибриден подход, комбиниращ статични и адаптивни механизми.

2.2.2. Сравнение, предизвикателства и ползи от различните подходи

Статичните методи за персонализация се характеризират с ограничена функционалност, тъй като разчитат основно на предварително зададени потребителски настройки и не осигуряват динамична адаптация на съдържанието. Поради това тяхното приложение е оправдано предимно като средство за преодоляване на проблема със студения старт или като допълващ източник на информация при оценка на потенциалния интерес към даден ресурс.

Анализът на потребителското поведение в уеб среда позволява предоставяне на адаптивно съдържание, но е свързан със значителни предизвикателства, произтичащи от събирането, почистването и обработката на регистрите за достъп и взаимодействие. Подходът е приложим основно в уеб базирани системи, включително дигитални библиотеки, при наличие на надеждни регистри за потребителско поведение, които могат да бъдат интегрирани с алгоритми за персонализирани препоръки. Допълнително предизвикателство произтича от динамичния характер на потребителските интереси, които се изменят във времето, което затруднява изграждането на устойчиви модели на поведение и намалява точността на дългосрочните прогнози за потребителските предпочитания.

Адаптивните методи като цяло демонстрират по-ниска ефективност в началните етапи на експлоатация поради липса на достатъчно данни, докато при натрупване на големи обеми от информация възникват изисквания за висока изчислителна ефективност и бърза генерация на препоръки в реално време [27]. Допълнително ограничение представлява оскъдността на потребителските оценки, което затруднява разграничаването между положителни и отрицателни примери. Въпреки това редица изследвания показват, че съвместното филтриране постига висока ефективност при наличие на достатъчно богати поведенчески данни и възможност за идентифициране на групи потребители със сходни интереси.

Както е отбелязано в [91], използването на единен подход за препоръчване рядко осигурява устойчиви резултати за всички типове данни. Практически реализации, като ShareTEC [29], демонстрират, че хибридният подход представлява по-ефективно решение, комбиниращо силните страни на статичните и адаптивните методи.

В този контекст статичните методи могат да бъдат използвани в началната фаза на системата за осигуряване на резервен подход на персонализация, основано на потребителски предпочитания, ключови думи, предварително дефинирани характеристики или глобална популярност на ресурси. Критично условие за това е наличието на богато описани ресурси с адекватни метаданни и семантични атрибути.

Паралелно с това е необходимо систематично съхранение и обработване на данни за потребителското поведение и оценките на ресурсите, така че след натрупване на достатъчен обем информация да може да се приложи съвместното филтриране като основен подход за персонализация. В този смисъл статичните методи изпълняват поддържаща роля, докато адаптивните подходи - включващи анализ на поведението, препоръчващи алгоритми и методи за класификация и клъстеризация - осигуряват по-висока степен на интелигентност, гъвкавост и устойчивост на персонализираните информационни системи.

Обобщение на разгледаните подходи за персонализация, техните предимства, предизвикателства и области на приложение е представено в таблица 2.

Таблица 2. Сравнение между подходите и методите за персонализация

<i>Подход / Метод</i>	<i>Предимства</i>	<i>Предизвикателства</i>	<i>Приложение</i>
<i>Традиционни подходи</i>			
<i>Анкети, ключови думи и статична персонализация</i>	Лесна за реализация; не изисква сложни алгоритми; потребителят има пряк контрол върху предпочитанията.	Не се адаптира към промени в поведението; изисква активно участие от потребителя; ниска точност при големи обеми данни.	Подходяща за начално профилиране, базови системи за съдържание и малки уеб портали.
<i>Профили и анализ на потребителското поведение в уеб среда</i>	Позволява автоматично извличане на поведенчески модели; работи с реални данни; подпомага динамична адаптация на съдържанието.	Изисква големи обеми данни и изчислителни ресурси; потенциални рискове за поверителността; чувствителен към динамично изменящи се интереси.	Електронна търговия, образователни платформи, социални мрежи, анализ на уеб трафик, където има налични данни от регистрите на достъп.
<i>Методи за филтриране</i>			
<i>Филтриране, базирано на съдържание</i>	Осигурява персонализация на база характеристики на обектите; не зависи от други потребители; препоръките са обясними.	Ограничено разнообразие на резултатите („ефект на филтърния балон“); изисква качествено описание на съдържанието.	Препоръки на продукти, новини, учебни материали, филми, музика.
<i>Съвместно филтриране, базирано на потребители</i>	Използва сходство между потребители; предлага разнообразни и контекстуално подходящи резултати;	Проблем с „нови“ потребители и обекти (проблем на студения старт); намалена точност при	Онлайн магазини, стрийминг платформи, социални приложения.

<i>Подход / Метод</i>	<i>Предимства</i>	<i>Предизвикателства</i>	<i>Приложение</i>
	обучаваща се архитектура.	разнородни интереси; високи изисквания към данните.	
<i>Съвместно филтриране, базирано на елементи</i>	Изчислително по-ефективен от филтрирането базираното на потребители при големи системи; устойчив при динамична промяна на потребителите; по-добра мащабируемост.	Зависим от богатството и качеството на историята на взаимодействия; трудности при нови обекти (студен старт).	Системи за препоръки на продукти, статии, видеа и други дигитални ресурси.
<i>Хибридни модели за филтриране</i>	Комбинират предимствата на съдържателно и съвместно филтриране; по-висока точност; намаляват ефекта от „студения старт“.	Повишена сложност и нужда от синхронизация между различни модели; трудна интерпретация.	Адаптивни образователни системи, персонализирани портали, мултимедийни препоръки.
<i>Алгоритми за машинно обучение</i>			
<i>К-средните</i>	Лесен за разбиране и имплементация; ефективен при умерени до големи обеми данни; осигурява бързо групиране.	Изисква предварително задаване на броя клъстери; чувствителен към начални стойности; не отчита припокриване между групи.	Сегментация на потребители, анализ на поведение, откриване на типични модели.

<i>Подход / Метод</i>	<i>Предимства</i>	<i>Предизвикателства</i>	<i>Приложение</i>
<i>Размитите к-средни</i>	Позволява „мека“ принадлежност на обектите към няколко клъстера; по-гъвкав при размити граници; по-точен при сложни данни.	Повишена изчислителна сложност; необходимост от параметър за степен на принадлежност; чувствителен към начални условия.	Анализ на поведение, адаптивни системи, контекстно зависима персонализация.
<i>Наивен Бейсов</i>	Висока ефективност при класификация; ниски изчислителни изисквания; лесен за обучение и внедряване.	Предполага независимост на признаците, което често не е реалистично; чувствителен към липсващи данни.	Класификация на съдържание, анализ на обратна връзка, прогнозиране на интереси.
<i>Машина на опорните вектори</i>	Висока точност при нелинейни зависимости; добра генерализация върху нови данни; стабилен при малки набори.	Висока изчислителна сложност; труден избор на ядро и параметри; слаба обяснимост.	Анализ на текст, изображения, предпочитания и модели в адаптивни системи.
<i>Обучение с частичен надзор</i>	Комбиниращ обучение с учител и без учител; използва големи количества неклаифицирани данни; подходящ при ограничени ресурси.	Чувствителен към грешки в частично етикетирани данни; изисква баланс между двата типа обучение.	Системи за препоръки, предсказване на поведение, образователни и социални приложения.

2.2.3. Представяния на текст и мерки за близост при персонализация

В контекста на персонализация, ключов е изборът на представяне на текста и мярката за близост. Настоящият раздел синтезира утвърдените подходи за векторизация

(бройчно представяне, хеширащо кодиране, претеглено представяне и вграждане на думи) и съответните мерки за сходство (косинус и Жакар).

Бройчно представяне (CountVectorizer)

Машините не могат да разбират символи и думи, затова, когато се работи с текстови данни, трябва да бъдат представени в числа, за да бъдат разбрани от машината [92]. Бройчно представяне реализира такова преобразуване, като изгражда речник от уникални термини и за всеки документ изчислява колко пъти се среща всеки термин; на фигура 7 е показан пример за тази трансформация. При този подход се изгражда речник от уникалните термини в корпуса с индекс за всеки термин, а после се конструира и разрежена матрица „документи x термини“ (известна и като матрица „термини-документи“), в която клетката (i, j) съдържа броя срещания на термин j в документ i . На практика се използват разрежени формати, тъй като всеки отделен документ съдържа малка част от речника и повечето клетки са нули. За улавяне на локални зависимости могат да се включат n -грамни признаци (напр. (1,2) за думи и биграми). Бройчното представяне е прозрачно и бързо за изграждане и осигурява силна базова линия, но третира термините като независими признаци и не улавя семантични връзки; поради това често се комбинира с намаляване на размерността преди клъстеризиране.

текст = ["Аз обичам и да чета книги и да гледам филми"]

Аз	обичам	и	да	чета	книги	гледам	филми
1	1	2	2	1	1	1	1

Фигура 7. Използване на CountVectorizer

Хеширащо кодиране (HashingVectorizer)

Хеширащо кодиране е техника за извличане на признаци, при която всеки токен се съпоставя чрез хешираща функция към индекс в предварително фиксирано пространство и така колекцията от документи се преобразува в разрежена матрица от честоти [93]. За разлика от бройчното представяне и претегленото, които поддържат явен речник и присвояват уникален индекс на всеки термин, при хеширането не се съхранява речник, което едновременно е предимство и ограничение [93]. Предимството е свързано

с ефективност на паметта и мащабируемост при много големи корпуси или потокови данни: размерността е предварително зададена и не расте с речника, което прави метода подходящ за поетапно/онлайн обучение [93]. Ограниченията произтичат от невъзможността за обратна трансформация (индекс \rightarrow термин) и от риска за хеш-сблъсъци, при който различни термини попадат в една и съща колона; последният ефект се смекчава чрез достатъчно голямо хеш-пространство, но не се елиминира напълно [93]. В този смисъл хеширащото кодиране и броячното представяне изпълняват сходна функция - „преобразуват колекция от текстове в матрица от честоти“ - но първият заменя явния речник с детерминистично хеширане и фиксирана размерност [93].

Претеглено представяне (TfidfVectorizer) и мярката TF-IDF

TF-IDF, съкратено от Term Frequency-Inverse Document Frequency (честота на термините и обратна честота на документите), е статистическа мярка, която оценява значимостта на термин за конкретен документ в рамките на колекцията [94]. Тя се получава като произведение от честотата на термина в документа и обратната документна честота в корпуса [95]:

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

Където двата показателя - честотата на дума в даден документ и обратната документална честота на думата в набор от документи - се изчисляват със следните формули:

$$tf(t, d) = \log(1 + freq(t, d))$$

$$idf(t, D) = \log\left(\frac{N}{count(d \in D, t \in d)}\right)$$

Компонентът TF отчита относителната важност на термина в рамките на дадения документ (думи с по-висока честота са по-значими) [96], а IDF потиска термините, които са чести в много документи (напр. общи функционални думи), като по този начин намалява тяхната тежест [95]. Така често срещаните в целия корпус думи получават ниски стойности, а редките, но характерни за конкретен документ - високи; TF-IDF може да се разглежда като мярка за „уникалност“ на термина спрямо документа и колекцията [95], [96].

Претеглено представяне прилага тази мярка върху цялата колекция от документи, като конструира матрица „документ x термин“, в която стойностите са TF-IDF теглата за

всеки термин във всеки документ [97]. Спрямо простото броене TF-IDF по-ефективно потиска тривиалните често срещани думи и често подобрява тематичното разграничаване, което е полезно както за извличане на информация, така и за изчисляване на сходство между документи [97].

Вграждане на думи (Embeddings)

Методите за вграждане представляват усъвършенстван подход за числово представяне на текст, при който се запазва семантична и смислова структура на езика. За разлика от класическите разредени представяния, които третира думите като независими признаци, моделите за вграждане обучават плътни векторни представяния въз основа на контекста, като по този начин улавят семантични, синтактични и дори прагматични зависимости [98], [99].

Класическите статични вграждания (напр. Word2Vec, GloVe) присвояват на всяка лексема едно и също векторно представяне независимо от контекста, докато контекстуалните вграждания, реализирани чрез трансформър архитектури (напр. BERT, RoBERTa, Sentence-BERT), генерират различни вектори за една и съща дума според употребата ѝ в конкретното изречение [99], [100]. Това води до по-добро улавяне на семантиката и синонимията на думите [101]. Практическо предимство е, че представянията са нискомерни и плътни спрямо традиционните разредени матрици, като същевременно съхраняват семантичната близост между понятията [101].

Наред с предимствата, подходът има и ограничения. Обучението и прилагането на съвременни модели изисква значителни изчислителни ресурси и големи корпуси, а качеството на данните е критично за избягване на пристрастия в представянията. Допълнително, високата абстракция и ограничената прозрачност на получените векторни пространства затрудняват интерпретацията и проследимостта на решенията, което е важно в приложения с изисквания за обяснимост [102]. Въпреки тези предизвикателства, вгражданията се утвърждават като стандартен инструмент в съвременната обработка на естествен език, демонстрирайки значително по-висока семантична изразителност спрямо класическите методи за векторизация [102].

Сравнение на методите за векторизация

Броячното представяне предоставя най-простото и прозрачно представяне, което го прави подходящ при сценарии, в които е необходима лесна интерпретация на признаците. Хеширащото кодиране, от своя страна, реализира същата обща цел -

преобразуване на текстовете в разрежена матрица от признаци - но го прави без поддръжка на явен речник и с предварително фиксирана размерност, което го прави особено ефективен при много големи или потокови данни; недостатък са възможните хеш-сблъсъци и невъзможността за обратна интерпретация на признаците [93]. Претегленото представяне надгражда този подход, като отчита разпределението на термините в целия корпус и понижава тежестта на често срещаните думи; в резултат се получава по-добро тематично разграничаване и по-подходящо представяне за задачи по извличане на информация и изчисляване на сходство [94], [95], [96], [97].

Методите за вграждане (embeddings) се отличават от горепосочените по това, че предоставят плътни и семантично обосновани представяния, в които близостта между векторите отразява близост на значенията [98], [99], [103], [104], [105], [102], [100], [101]. За разлика от първите три, които третираат думите като независими признаци и основно кодират честота, вгражданията могат да уловят контекст и синонимия и поради това са по-подходящи за задачи, при които е важно семантичната близост да бъде запазена (напр. показване на подобни текстове, препоръчване по съдържание, клъстеризиране на тематично близки документи). Ограничението им е свързано с по-висока изчислителна цена и по-ниска прозрачност, което има значение при приложения с изискване за обяснимост. В Таблица 3 са сравнени тези четири подхода като се описани предимствата и ограниченията на всеки от тях.

Таблица 3. Сравнение на основните методи на кодиране на текст

<i>Метод</i>	<i>Основна идея</i>	<i>Предимства</i>	<i>Ограничения</i>
<i>Броячно представяне</i>	Представя документ чрез честотата на всяка дума, загуба на информация за подредба и контекст.	Прост за имплементиране и разбиране; работи добре при малки до средни набори от данни.	Не улавя семантика, контекст; води до висока размерност
<i>Претеглено представяне</i>	Разширява броячното представяне чрез претегляне: честа дума в документ +	По-добро от броячното представяне в много	Все още третира думите като независими

Метод	Основна идея	Предимства	Ограничения
	рядка дума в корпус води до по-голямо значение	случаи; по-лесно за интерпретация.	признаци; не улавя семантични връзки.
Статични вграждания на думи	Представя всяка дума като плътен вектор, така че думи с подобен контекст да са близки в пространството.	Улавя семантика и взаимовръзки между думи; редуцира размерността сравнително.	Векторите са фиксирани за всяка дума (не отчита контекст); моделирането е по-ограничено.
Контекстуални вграждания	Векторите за думата се променят в зависимост от контекста, в който се намира думата.	Най-висока семантична и синтактична изразителност; сила при сложни задачи.	По-високи изчислителни изисквания; трудна интерпретация.

Метрики за близост

Косинусова близост - Мярка за сходство между два ненулеви вектора, дефинирана като косинус на ъгъла помежду им: $\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$. Стойностите са в $[-1, 1]$, а при неотрицателни представяния (напр. TF-IDF) в $[0, 1]$ [106], [107]. Тя е инвариантна спрямо мащаба (чувствителна към посока, не към дължина), поради което е стандарт за текстови вектори и вграждания [106]. Подходяща е, когато значението се кодира в ориентацията на вектори, а не в абсолютната им норма [106].

Коефициент на Жакар - Мярка за сходство между два множества A и B, дефинирана като отношение на сечението към обединението: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, със стойности в $[0, 1]$ [107]. Широко използвана при двоични признаци и „торбички“ от обекти (напр. множества от именувани същности, ключови термини) [107]. Има и претеглени обобщения за мултимножества [107]. Подходяща е, когато интерес представлява споделеното покритие (общите елементи) спрямо общия обем [107].

Обобщено, различните представяния на текст и мерки за близост предлагат различен баланс между изчислителна цена, обяснимост и семантична изразителност.

В следващата глава се представя методологична рамка за емпирично сравнение на приложимите подходи в контекста на дигитални библиотеки.

2.2.4. Преглед на съвременни изследвания и използвани методи за персонализация в дигитални библиотеки

Въведение и обхват на прегледа

Настоящият литературен преглед разглежда и обобщава съвременните научни изследвания, посветени на персонализацията, потребителското преживяване, системите за препоръки, етичните измерения на използваните технологии, както и на приложението на изкуствения интелект и машинното обучение в дигитални библиотеки. Целта е от една страна е да се изгради цялостно разбиране за развитието на подходите за персонализация и тяхното въздействие върху начина, по който потребителите взаимодействат с информационните ресурси и услуги и от друга страна да се идентифицират проблемни области и предизвикателства, които да бъдат частично разрешени или облекчени.

В рамките на анализа са проучени **113** научни публикации, подбрани след филтрация на **475** първоначални източници, публикувани в периода 2018 - 2025 г. Прегледът включва емпирични, теоретични и приложни изследвания, фокусирани върху персонализация, анализ на потребителското поведение, алгоритми за препоръки, етични и социални аспекти на персонализацията, както и интеграцията на системи, базирани на изкуствен интелект и машинно обучение, в съвременните дигитални библиотеки.

Чрез анализа на съществуващите тенденции, предизвикателства и етични измерения, изследването цели да идентифицира основните пропуски в научните разработки и практическите приложения и да предложи концептуален модел, способен да смекчи част от установените ограничения. По този начин се създават предпоставки за разработване на по-ефективни, прозрачни и ориентирани към потребителя системи за персонализация, които използват изкуствен интелект и машинно обучение за подобряване на достъпа до ресурси и за обогатяване на потребителското преживяване в дигиталните библиотеки.

В съответствие с общата цел на изследването, настоящият преглед си поставя следните конкретни задачи:

- Да обобщи и оцени публикациите от последните години относно стратегиите за персонализация и тяхното въздействие върху потребителското преживяване в дигитални библиотеки;

- Да извърши критичен сравнителен анализ на алгоритмите за препоръки, базирани на изкуствен интелект и машинно обучение;
- Да анализира емпирични изследвания и практически примери, илюстриращи ефективността и ограниченията на решенията за персонализация, задвижвани от изкуствен интелект и машинното обучение;
- Да открие съществуващите изследователски пропуски и да предложи насоки за разработването на адаптивни и етично устойчиви модели за персонализация в дигитални библиотеки.

Концептуална рамка на изследванията

Персонализацията в дигиталните библиотеки се очертава като приоритетно изследователско направление, мотивирано от процесите на мащабна дигитализация, водещи до информационно пренасищане, както и от необходимостта от подобряване на потребителското преживяване и ефективността на откриване на ресурси в големи цифрови колекции [15], [22]. През последното десетилетие развитието на дигиталните библиотеки се характеризира с интегрирането на усъвършенствани системи за препоръки и анализ на потребителското поведение, като тенденцията се измества от традиционно филтриране на съдържанието към комплексна персонализация, базирана на изкуствен интелект и алгоритми от машинното обучение [108], [109], [110]. Този напредък подчертава социалната и практическата значимост на персонализираните подходи, които повишават удовлетвореността и ангажираността на потребителите; емпирични проучвания отчитат подобрения в точността на препоръките и задържането на потребителите, надхвърлящи 80% в някои системи [111], [112]. Практическите приложения обхващат академични, обществени и интелигентни библиотеки, в които персонализирани услуги се разглеждат като стратегия за справяне с информационното претоварване и разнообразните потребителски нужди [113], [114].

Въпреки отчетения напредък, остават значителни предизвикателства при ефективното прилагане на персонализирани услуги. Сред най-съществените проблеми се открояват явленията на „студен старт“, оскъдността на данни, въпросите, свързани с поверителността, и алгоритмичните пристрастия, които ограничават качеството на препоръките и доверието на потребителите [15], [115]. Отбелязват се и пропуски в разглеждането на етичните съображения и мащабируемостта, като реалните примери за прозрачни и етично устойчиви приложения на изкуствен интелект остават ограничени

[15], [116]. Необходим е балансиран подход между максимизиране на точността и ресурсите и времето за генериране на препоръките [114], [117].

В рамките на прегледа персонализацията, потребителското преживяване и системите за препоръки се разглеждат като тясно свързани направления в развитието на дигиталните библиотеки [22], [118], [119]. Персонализацията се дефинира като процес на адаптиране на съдържанието и услугите към индивидуалните предпочитания на потребителите чрез анализ на потребителското поведение и прилагане на техники от областта на изкуствения интелект и машинното обучение [22], [113]. В контекста на персонализираните системи потребителското преживяване се оценява чрез показатели като удовлетвореност, ангажираност и ефективност [120], [121]. Системите за препоръки прилагат персонализацията чрез алгоритми като съвместно филтриране, филтриране на базата на съдържание и хибридни модели за предоставяне на релевантни и контекстуално подходящи ресурси [117], [118].

Прегледът на публикациите от последните години показва няколко основни направления, сред които модели за препоръчване, базирани на алгоритми и потребителско поведение, адаптивни методи на машинно обучение, използване на семантични и контекстуални фактори, както и изследвания, свързани с потребителското преживяване и етичните въпроси. Тези направления водят до развитието на интелигентни системи, които не само предлагат подходящо съдържание, но и се адаптират към промените в поведението и нуждите на потребителите.

Анализът на съвременните публикации показва, че повече от шестдесет изследвания (от разгледаните 113) демонстрират значително повишение на удовлетвореността и ангажираността посредством персонализация, базирана на изкуствен интелект; особено високи резултати се отчитат при хибридни модели, подсилващо обучение и дълбоки невронни мрежи [22], [109], [111]. Подчертава се ключовата роля на динамичното моделиране на интересите и поведенческия анализ за ефективно адаптиране на препоръките [15], [122], [123], както и приносът на контекстуални и културни фактори за уместност и лоялност, включително специализирани подходи по настроение, възраст или академична роля [112], [124], [125], [126].

От гледна точка на алгоритмите, редица проучвания докладват висока точност, прецизност и стабилност при използване на графови невронни мрежи, механизми за внимание и хибридно филтриране [127], [128], [129], [130]. Хибридните подходи,

комбиниращи съвместно и съдържателно филтриране, превъзхождат традиционните решения, особено при „студен старт“ и оскъдни данни [115], [118], [131]. Нови техники като извличане на асоциативни правила и хетерогенни мрежови вграждания подобряват мащабируемостта и качеството на препоръките [109], [132], а валидации върху реални набори от данни потвърждават устойчивостта им в различни библиотечни контексти [111], [133], [134].

Паралелно, около тридесет публикации акцентират върху етичните и нормативните измерения на персонализацията, поставяйки акцент върху прозрачност, защита на лични данни и отговорно внедряване на изкуствения интелект [15], [22], [113]. Като ключови проблеми се открояват алгоритмичната пристрастност, информираното съгласие и справедливият достъп до информация [114], [135]. Същевременно значителна част от технологично ориентираните изследвания не включват задълбочен етичен анализ - съществен пропуск в литературата [109], [136].

Наред с теоретичните разработки, множество публикации представят успешни приложения в академични и публични библиотеки, показващи по-добро приемане от потребителите, мащабируемост и позитивна обратна връзка [114], [137], [138]. Разработките обаче често отбелязват ограничения, свързани с инфраструктура, финанси и поверителност [110], [139]. Оценки чрез потребителски проучвания и системни симулации потвърждават подобро преживяване и по-висока релевантност на препоръките [121], а иновативни решения с виртуални асистенти и потапящи (имерсивни) среди разширяват възможностите на библиотечните услуги [140], [141].

Таблица 4 синтезира разгледаните в литературата подходи за персонализацията, като очертава ключовите им предимства, ограничения и типични контексти на приложение за периода 2018-2025 г.

Таблица 4. Подходи за персонализацията в дигиталните библиотеки: силни и слаби страни

<i>Аспект</i>	<i>Силни Страни</i>	<i>Слаби Страни</i>
<i>Алгоритмични подходи и ефективност</i>	Литературата представя широка гама от алгоритми за изкуствен интелект и машинно обучение, включително модели за дълбоко	Много изследвания използват експериментални набори от данни и симулации вместо широкомащабни реални

<i>Аспект</i>	<i>Силни Страни</i>	<i>Слаби Страни</i>
	<p>обучение, подсилващо обучение, графични невронни мрежи и хибридни системи за препоръки, които са демонстрирали подобрена точност, адаптивност и мащабируемост в персонализираните препоръки [109], [115], [127], [142]. Тези методи подобряват точността на препоръките и подпомагат ангажираността на потребителите, като отчитат връзките между потребители и елементи и тяхната промяна във времето [122], [143].</p>	<p>приложения, което ограничава приложимостта на получените резултати [109], [131]. Освен това, някои алгоритми, макар и точни, изискват значителни изчислителни ресурси, което може да затрудни внедряването им в дигитални библиотеки с ограничени ресурси [129], [130]. Проблемът с „студения старт“ остава постоянна предизвикателство въпреки хибридните подходи [115].</p>
<i>Анализ и профилиране на поведението на потребителите</i>	<p>Изследванията подчертават критичната роля на анализа и профилирането на поведението на потребителите при персонализирането на услугите, като моделите използват регистрите на взаимодействия със системата и семантичната класификация за създаване на динамични потребителски профили [22], [123], [144]. Интеграцията на механизми за внимание допълнително усъвършенства разбирането за променящите се интереси на потребителите [111], [143].</p>	<p>Изследванията върху разнообразието и приобщаването на потребителските профили са ограничени, като недостатъчно внимание се отделя на слабо представените групи потребители и на различните културни контексти [112]. Макар рисковете, свързани със събирането на потребителски данни, да са добре известни, те рядко се разглеждат систематично в процеса на системен дизайн [140].</p>
<i>Етични съображения и съображения,</i>	<p>Няколко проучвания признават етичните предизвикателства, включително поверителността на</p>	<p>Въпреки отчитането на това като съществен пропуск, практическите решения за</p>

<i>Аспект</i>	<i>Силни Страни</i>	<i>Слаби Страни</i>
<i>свързани с поверителността</i>	данните, алгоритмичната пристрастност и прозрачността, като подчертават необходимостта от внедряване на изкуствен интелект и механизми за съгласие на потребителите [15], [113], [114]. Публикациите изтъкват важността на балансирането на ползите от персонализацията със защитата на правата на потребителите [114].	намаляване на етичните рискове са недостатъчно развити, като има малко примери за прилагане на надеждни техники за защита на личните данни или стратегии за намаляване на пристрастността [145]. В литературата липсват изчерпателни рамки за текуща етична оценка и изграждане на доверие у потребителите в персонализацията, основана на изкуствен интелект [116].
<i>Казуси и практически приложения</i>	Казусите илюстрират успешната персонализация, задвижвана от изкуствен интелект, в различни дигитални библиотеки, като показват подобрения в удовлетвореността на потребителите и използването на ресурсите [112], [146]. Някои системи включват обратна връзка от потребителите и взаимодействие в реално време, което подобрява адаптивността и ангажираността на потребителите [121], [147].	Много от казусите са ограничени по мащаб или продължителност, често включват малки групи потребители или контролирани среди, което ограничава познанията за дългосрочната ефективност и мащабируемост [121]. В литературата се наблюдава ограничен брой изследвания, посветени на интеграцията в съществуващите библиотечни системи и подготовката на персонала [114], [139].
<i>Интегриране на изкуствен интелект и машинното обучение за</i>	Доказано е, че приложенията на изкуствения интелект и машинното обучение значително подобряват потребителското преживяване, като позволяват	Въпреки технологичните постижения дизайнът на потребителския интерфейс и съобщенията за използваемост понякога са

<i>Аспект</i>	<i>Силни Страни</i>	<i>Слаби Страни</i>
<i>подобряване на потребителското преживяване</i>	персонализирани препоръки, семантично търсене и адаптивни интерфейси, което допринася за по-висока удовлетвореност и лоялност на потребителите [110], [120], [148]. Използването на обработка на естествен език и подходи за внимание подобрява уместността и разнообразието на препоръките [129], [130], [149].	второстепенни, което води до потенциални несъответствия между алгоритмичните резултати и очакванията на потребителите [120]. Освен това рискът от прекомерна зависимост от изкуствен интелект може намали човешкото взаимодействие и критичната оценка от страна на потребителите [150], [151].
<i>Мащабируемост и производителност на системата</i>	Хибридни модели, комбиниращи съвместно филтриране, филтриране на базата на съдържание и техники за клъстеризиране, демонстрират подобрена мащабируемост и разнообразие на препоръките, като отговарят на някои ограничения на традиционните методи [115], [131], [137]. За практическо приложение са предложени ефективни алгоритми с ниски изчислителни разходи [109].	Мащабируемостта, устойчивостта при студен старт и балансът между изчислителната сложност и качеството на препоръките остават основни предизвикателства. [129], [130], [131], [132], [145]

Таблица 5 и фигура 8 показва честотата на използваните техники в анализираниите (113) публикации, като отразява доминиращите практики и тенденции за периода 2018-2025 г.

Таблица 5. Честота на използваните техники за персонализация в анализираните публикации

<i>Тема</i>	<i>Брой</i>	<i>Описание</i>
<i>Техники от изкуствен интелект и машинно обучение</i>	65/113	Съвременни методи от изкуствения интелект и машинното обучение, включително съвместно филтриране, дълбоко и подсилващо обучение, графични невронни мрежи и големи езикови модели, се прилагат за повишаване на точността на препоръките, преодоляване на проблемите на студения старт и разредеността, както и за моделиране на динамичните потребителски интереси в дигитални библиотеки. Тези подходи допринасят за по-ефективна персонализация, откриване на ресурси и мащабируемост на системите. [109], [127], [128], [131], [152], [153].
<i>Анализ на поведението на потребителите и профилиране на потребителите</i>	52/113	Анализът на потребителското поведение чрез извличане на данни, клъстеризация и моделиране на потребителски профили е ключов за разработването на ефективни стратегии за персонализация. Техники като потребителско моделиране, механизми за насочване на вниманието и семантична класификация подпомагат адаптацията към динамичните предпочитания, повишават релевантността на препоръките и подобряват потребителската удовлетвореност. [15], [22], [119], [144]
<i>Хибридни и контекстно-ориентирани модели за препоръки</i>	38/113	Хибридните препоръчващи системи, комбиниращи съвместно и съдържателно филтриране, извличане на асоциативни правила и онтологична интеграция, смекчават ограничения като студен старт и разреденост на данните. Контекстно-ориентираните подходи допълнително повишават степента на персонализация чрез отчитане на потребителския контекст и многоатрибутни зависимости. [118], [154], [155], [156].
<i>Етични предизвикателства и опасения за поверителността</i>	34/113	Изследванията акцентират върху етични аспекти като защита на личните данни, алгоритмична пристрастност, прозрачност, справедливост и информирано съгласие при персонализираните услуги в дигиталните библиотеки. Тези

<i>Тема</i>	<i>Брой</i>	<i>Описание</i>
		фактори оказват съществено влияние върху проектирането на системите с цел осигуряване на отговорно използване на изкуствения интелект, равнопоставен достъп и потребителско доверие. [15], [113], [114], [116]
<i>Практически приложения и казуси</i>	31/113	Множество изследвания демонстрират практическото приложение на персонализация, базирана на изкуствен интелект, в дигитални библиотеки, отчитайки повишена потребителска ангажираност, по-ефективно използване на ресурсите и по-висока удовлетвореност. Представените казуси обхващат университетски и публични библиотеки и цифрови хранилища, като илюстрират разнообразни технологични интеграции и постигнати резултати. [109], [140], [146].
<i>Приложения на дълбокото обучение и невронните мрежи</i>	29/113	Подходите за дълбоко обучение, механизми за внимание и дълбоки невронни мрежи, се използват все по-често за моделиране на временното поведение на потребителите, подобряване на семантичното разбиране и повишаване на ефективността на препоръки в дигитални библиотеки [142], [152], [157], [158], [159].
<i>Усъвършенствана обработка на естествен език (NLP) и семантични техники</i>	22/113	Техниките за обработка на естествен език и семантично моделиране подпомагат анализа на съдържанието и откриването на потребителските интереси, като подобряват точността на препоръките и търсенето [133], [144], [149], [160].
<i>Решаване на проблема с „студения старт“ и оскъдността на данни</i>	20/113	Стратегиите за смекчаване на предизвикателствата, свързани със „студения старт“ и оскъдността на данни, включват клъстеризация, размита логика, библиометрични данни и хибридни подходи за филтрация. Тези методи улесняват ефективното моделиране на потребителите и препоръките дори при ограничени потребителски данни [115], [161].
<i>Интегриране на изкуствен</i>	17/113	Решенията, базирани на изкуствен интелект, разширяват възможностите за персонализация чрез предоставяне на по-

Тема	Брой	Описание
интелект за приобщаващи и адаптивни потребителски преживявания		достъпни и приобщаващи библиотечни услуги, включително виртуални асистенти, инструменти за достъпност и адаптивни интерфейси, като по този начин повишават ангажираността и удовлетвореността на потребителските групи [113], [139].
Рамки за оценка и проучвания на потребителското преживяване	15/113	Проучванията подчертават важноста на внедряването на методи за препоръки в рамките на потребителското преживяване, оценяването на използваемостта на системи за удовлетвореността на потребителите и контекстуалните влияния върху ефективността на персонализацията [121], [124], [162].



Фигура 8. Честота на използваните техники за персонализация в анализираните публикации

Обобщено, прегледът на литературата показва значителен напредък в персонализацията на дигиталните библиотеки, постигнат чрез използването на хибридни и дълбоки модели. В същото време остават предизвикателства, свързани със студения старт, ограничената наличност на данни, прозрачността и етичните аспекти. Въпреки подобренията в точността на препоръките и потребителската ангажираност, липсват

утвърдени рамки, които едновременно да отчитат мащабируемостта, справедливостта и контекстовата уместност в реални условия. На тази основа следващият подраздел идентифицира основните изследователски пропуски и предлага насоки за тяхното преодоляване.

Пропуски в изследванията

На фона на постигнатите резултати в областта на персонализацията в дигиталните библиотеки, настоящият раздел насочва вниманието към основните пропуски в съществуващите изследвания.

- Множество публикации посочват липсата на устойчиви етични рамки, ограничената мащабируемост на системите за персонализация и необходимостта от интегриране на разнообразни източници на данни [15], [116], [117].
- Недостатъчно са изследвани дългосрочните ефекти на персонализацията, основана на изкуствен интелект, върху потребителското доверие и поведение [116], [163].
- Някои изследвания изтъкват ролята на сътрудничеството между различни научни области и отворените решения за постигане на по-достъпна персонализация [113].
- Нови направления като интегриране на виртуални среди, разширена реалност и усъвършенствана обработка на естествен език остават слабо проучени и методологично недоразвити [140], [160].

Таблица 6 обобщава основните области на ограничения, изведени въз основа на литературния преглед.

Таблица 6. Основни области на ограничения в литературата

<i>Област на ограничение</i>	<i>Описание на ограничението</i>
<i>Ограничен фокус върху етиката и поверителността</i>	Много изследвания признават значението на поверителността, етичните въпроси и алгоритмичната пристрастност, но предлагат ограничени емпирични доказателства или решения, което ограничава

	изводите относно отговорната персонализация и разбирането на доверието и справедливостта в системите, използващи изкуствен интелект [15], [113], [114]
<i>Проблеми със „студения старт“ и оскъдността на данни</i>	Някои системи за препоръки срещат затруднения, свързани със студения старт и ограничените данни за взаимодействие между потребители и елементи, което намалява ефективността им при нови потребители или елементи и ограничава обобщаемостта и устойчивостта на подходите за персонализация [115], [131].
<i>Малки или ограничени потребителски извадки</i>	Някои емпирични оценки се базират на ограничени или локални потребителски извадки, което намалява обобщаемостта на резултатите и външната им валидност в по-широки дигитални библиотеки [112], [121], [150].
<i>Липса на мултимодална интеграция на данни</i>	Преобладаващото използване на текстови данни пренебрегва мултимодалната информация, като изображения, видео и аудио, което ограничава дълбочината на моделирането на потребителите и точността на препоръките [145].
<i>Недостатъчно внедряване в реалния свят</i>	Много от предложените модели и алгоритми се тестват предимно на офлайн бази от данни или симулации, като липсват обширни приложения в реалния свят и проучвания на взаимодействието с потребителите, което ограничава разбирането за практическата използваемост и дългосрочната ефективност [127], [146].
<i>Ограничения на показателите за оценка</i>	Често използваните показатели като точност и прецизност не отразяват напълно удовлетвореността на потребителите или дългосрочното им ангажиране, което води до непълна оценка на ефективността на системата за препоръки и потребителското преживяване [120], [145], [153].
<i>Мащабируемост и изчислителни ограничения</i>	Някои усъвършенствани модели на изкуствен интелект и машинно обучение срещат ограничения, свързани с мащабируемостта и изчислителната ефективност, което ограничава приложимостта им в

<i>Област на ограничение</i>	<i>Описание на ограничението</i>
	мощни дигитални библиотеки и влияе върху бързодействието на системата и потребителското преживяване [109], [129], [130].
<i>Ограничен фокус върху разнообразните нужди на потребителите</i>	Съществуващите подходи за персонализация често не отчитат в достатъчна степен разнообразието на потребителите, включително културните, езиковите и потребностите, свързани с достъпността, което ограничава равнопоставения достъп и обхвата на услугите в дигиталните библиотеки [110], [113].
<i>Акцент върху алгоритмичната точност</i>	В литературата често се наблюдава приоритизиране на алгоритмичната точност за сметка на обяснимостта, прозрачността и потребителския контрол, което може да намали доверието и приемането на персонализираните препоръки в дигиталните библиотеки [114], [119], [142].

Таблица 7 обобщава основните пропуски, установени в литературния преглед, и предлага насоки за бъдещи изследвания.

Таблица 7. Пропуски и насоки за бъдещи изследвания и подобрения

<i>Област на пропуски</i>	<i>Описание</i>	<i>Бъдещи насоки за изследвания</i>	<i>Обосновка</i>
<i>Етични рамки и защита на личните данни</i>	Много алгоритми за персонализация не разполагат с изчерпателни етични рамки и надеждни механизми за защита на личните данни в дигиталните библиотеки.	Разработване и прилагане на стандартизирани етични насоки и техники за защита на личните данни, адаптирани към системите за персонализация в	Разрешаването на въпросите, свързани с поверителността и етиката, е ключово за поддържането на потребителското доверие и спазването на нормативните изисквания, но съществуващите изследвания предлагат

<i>Област на пропуски</i>	<i>Описание</i>	<i>Бъдещи насоки за изследвания</i>	<i>Обосновка</i>
		дигиталните библиотеки.	ограничени практически решения [15], [116].
<i>Продължителни и мащабни емпирични изследвания</i>	Липсват дългосрочни и широкомащабни емпирични оценки на въздействието на персонализацията, базирана на изкуствен интелект, върху ангажираността на потребителите.	Провеждане на многоинституционални, дългосрочни проучвания за оценка на трайните ефекти от персонализацията върху удовлетвореността на потребителите, поведението им и използването на ресурсите в различни дигитални библиотеки.	Повечето съществуващи изследвания се основават на краткосрочни или мащабни казуси, което ограничава разбирането за дългосрочната ефективност и мащабируемост [112], [121].
<i>Предизвикателства, свързани със „студения старт“ и оскъдните данни</i>	Продължават да съществуват постоянни предизвикателства при ефективното решаване на проблемите с „студения старт“ и оскъдните данни за взаимодействието на потребителите в системите за препоръки.	Хибридни модели, интегриращи неточно клъстеризирането и мултимодални източници на данни, за да се подобри справянето със „студения старт“ и да се подобри точността на препоръките за нови потребители и елементи.	Въпреки хибридните подходи, студеният старт остава пречка, особено в сценарии с ограничени ресурси или нови потребители [115], [131].
<i>Обясними и прозрачни модели на</i>	Ограничени са изследванията, фокусирани върху	Разработване на обясними модели и потребителски	Обяснимостта е от съществено значение за етичното прилагане на

<i>Област на пропуски</i>	<i>Описание</i>	<i>Бъдещи насоки за изследвания</i>	<i>Обосновка</i>
<i>изкуствен интелект</i>	обяснимостта и прозрачността на алгоритмите, използващи изкуствен интелект и машинно обучение.	интерфейси, използващи методи от изкуствения интелект, с цел повишаване на доверието и отчетността.	изкуствен интелект и за приемането му от потребителите, но остава недостатъчно изследвана в съществуващите рамки за персонализация [114], [119].
<i>Интеграция на нововъзникващи технологии от изкуствен интелект</i>	Недостатъчно изследвани остават приложението на големи езикови модели, потапящите среди (VR/AR) и метавселенските технологии в персонализацията на дигиталните библиотеки.	В литературата се предлага изследване на дизайна, внедряването и въздействието върху потребителите на препоръчителни системи, базирани на големи езикови модели, виртуални асистенти, задвижвани от изкуствен интелект, и потапящи социални платформи в библиотечен контекст.	Новите технологии в областта на изкуствения интелект притежават значителен потенциал за трансформация, но в контекста на персонализацията на библиотеките липсват достатъчни емпирични изследвания и утвърдени практически рамки [128], [140], [141].
<i>Приобщаване и културна чувствителност</i>	В съществуващите изследвания се наблюдава недостатъчен фокус върху приобщаването, културната чувствителност и адаптирането на услугите към	Проектиране на адаптивни системи за персонализация, които включват културни, езикови и фактори за достъпност, валидирани чрез разнообразни	Ефективността на персонализацията варира в зависимост от културния и икономическия контекст, което налага необходимостта от индивидуален подход, за да се гарантира равнопоставено

<i>Област на пропуски</i>	<i>Описание</i>	<i>Бъдещи насоки за изследвания</i>	<i>Обосновка</i>
	потребителите от слабо представени и разнообразни групи	демографски проучвания на потребителите.	потребителско преживяване [112], [113], [161].
<i>Мащабируемост и изчислителна ефективност</i>	Проблемите с мащабируемостта продължават да се проявяват при обработката на големи и хетерогенни масиви от данни, както и при поддържането на препоръки в реално време в дигиталните библиотеки.	Разработване на леки, мащабируеми алгоритми, оптимизирани за среди с ограничени ресурси, като се използват крайни изчислителни системи и разпределени архитектури.	Много от съвременните модели изискват големи изчислителни ресурси, което ограничава внедряването им в библиотеки с ограничена инфраструктура [109], [129], [130], [132].
<i>Еволюция на потребителските интереси</i>	Настоящите анализи на поведението на потребителите често пренебрегват динамичните промени и вариативността на потребителските интереси във времето.	Необходимост от прилагане на методи за непрекъснато и контекстно ориентирано профилиране на потребителите, които отчитат времевата динамика, ситуационната осведоменост и множествеността на потребителските интереси.	Улавянето на динамично променящите се потребителски предпочитания е от съществено значение за постигането на точна персонализация, но остава недостатъчно изследвано [111], [164].
<i>Интердисциплинарно сътрудничество и</i>	Недостатъчни са интердисциплинарните изследвания и отворените платформи	Развитие на интердисциплинарни и отворени решения за персонализация.	Необходими са интердисциплинарни подходи и достъпни инструменти за

<i>Област на пропуски</i>	<i>Описание</i>	<i>Бъдещи насоки за изследвания</i>	<i>Обосновка</i>
<i>инструменти с отворен код</i>	за персонализация с изкуствен интелект в дигиталните библиотеки.		ускоряването на иновациите и етичното внедряване [113], [165].
<i>Показатели за оценка откъд точността</i>	Оценяването, базирано основно на точност, не отчита в достатъчна степен потребителските и етичните аспекти.	Прилагане на цялостни рамки за оценка с потребителски ориентирани показатели.	Настоящите показатели не отразяват адекватно реалното потребителско преживяване и етичните аспекти, което ограничава възможностите за усъвършенстване на системата [117], [145], [153].

Анализ и обобщение

Прегледаната литература показва висока степен на ангажираност с методи от изкуствения интелект и машинното обучение за усъвършенстване на персонализацията и препоръчващите системи в дигиталните библиотеки. Отчетен е съществен напредък в потребителското преживяване и ефективността на откриване на ресурси чрез прилагане на дълбоки невронни мрежи, подсилващо обучение, графови модели и различни хибридни схеми.

Наред с това анализът очертава няколко групи ограничения, които са особено релевантни за настоящото изследване:

- **Проблем на студения старт и оскъдни данни.** Значителна част от решенията разчитат на богати регистри на взаимодействията и демонстрират понижена ефективност при нови потребители, нови ресурси или неравномерно разпределени данни.
- **Машабируемост и изчислителна ефективност.** Използването на сложни модели, включително големи езикови модели, изисква значителни изчислителни ресурси и затруднява тяхното прилагане и обновяване в реални оперативни среди с нарастващи корпуси и динамични потребителски бази.

- **Обяснимост, прозрачност и етични изисквания.** Значителна част от алгоритмите остават трудни за интерпретация, което ограничава доверието, възможността за контрол и съответствието с етични и регулаторни рамки.
- **Ограничена хибридность и интеграция.** Често се използват изолирани източници на данни (само съдържателни или само поведенчески), без цялостна рамка за съгласувано комбиниране на текстови характеристики, метаданни, популярност и контекстна информация.

Наред с посочените ограничения, литературата разкрива и **методологични дефицити**, изразяващи се в липса на дългосрочни, мащабни и възпроизводими емпирични оценки в реална експлоатационна среда, както и в значителна хетерогенност на използваните методи, която затруднява сравнимостта и обобщаването на резултатите.

Идентифицираните ограничения мотивират разработването на обединяващи модели за персонализация, които интегрират поведенческо профилиране и съвместни препоръчващи подходи в единна архитектура и едновременно адресират мащабируемостта, изчислителната ефективност и изискванията за обяснимост, етична прозрачност и отчетност.

В този контекст следващите глави представят архитектура за адаптивна персонализация, които целят смекчаване на идентифицираните ограничения чрез съчетаване на съдържателни и поведенчески характеристики, контролирано и асинхронно прилагане на изчислително интензивни методи, както и повишена прозрачност на принципите за формиране на препоръките.

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

3.1. Въведение и обхват

В тази глава се представя методологичната основа и архитектурната организация на предложеното решение за персонализирано представяне на съдържание в дигитална библиотека под формата на „подобни текстове“ и персонализирани препоръки. Под персонализация се разбира интегрирането на съдържателни, поведенчески и семантични показатели, чрез които системата адаптира представянето на информационните ресурси към контекста на използване и установените предпочитания на конкретния потребител, като едновременно с това осигурява прозрачни основания за генерираните препоръки.

Подходът е **хибриден** и е насочен към **текстови ресурси** от тип **периодични многотематични издания на български език**. Той интегрира обработката на информационните **ресурси, регистрите на потребителските взаимодействия** и обогатените **метаданни** в единна архитектура, при която изчислително интензивните стъпки се изпълняват като отделни процеси извън оперативния поток.

Съществен фактор при проектирането на архитектурата са особеностите на използваните данни. Изследването е базирано върху дигитализиран фонд от периодични издания на Народна библиотека „Иван Вазов“ - Пловдив [166], който представлява хетерогенен корпус от текстови ресурси с различна тематика, времеви принадлежности, стилови характеристики и структура на съдържанието. За разлика от специализираните колекции, в които документите принадлежат към ясно дефиниран тематичен домейн, периодичните издания могат да обхващат обществени, политически, културни, исторически, икономически, образователни, научни теми и други теми в един и същ издание. В много случаи в рамките на един и същ брой на издание са включени публикации по напълно различни тематика, което значително усложнява процеса по автоматично идентифициране на тематична близост между документите и изграждането на надеждни препоръки.

Спецификата на тези данни поставя редица предизвикателства пред стандартните подходи за препоръчване и извличане на сходни документи. Многотематичният характер на публикациите означава, че близостта между два документа невинаги може да бъде определена единствено чрез сходството в използваната терминология. Документи със сходен речников състав могат да разглеждат напълно различни събития или контексти,

докато публикации с реална тематична връзка може да използват различен стил. Поради тази причина класически подходи за оценка на сходство, базирани основно на статистически характеристики на текста и метрики като косинусова близост, не винаги предоставят достатъчно надеждни резултати за този тип корпус. Ограниченията се проявяват особено ясно при функционалности от тип „подобни документи“, където е необходимо да бъдат откривани не само лексикално сходни, но и смислово свързани публикации.

Допълнително усложнение произтича от факта, че значителна част от връзките между документите се формират чрез общи личности, организации, институции, географски обекти, исторически събития и други контекстуални елементи, които невинаги могат да бъдат адекватно представени чрез традиционните модели за текстово сходство. Наред с това тематичното разпределение в колекцията е неравномерно, като определени направления са представени от значително повече публикации от други, което създава предпоставки за неравномерно генерирането на препоръки. Допълнително предизвикателство представлява и ограниченото количество потребителски взаимодействия спрямо общия обем на наличните документи, което води до висока степен на разреденост на данните и намалява ефективността на подходите, разчитащи единствено на поведенческа информация.

Поради тези характеристики архитектурата е проектирана така, че да комбинира съдържателни, семантични и поведенчески показатели при оценката на сходството между документите и при генерирането на персонализирани препоръки.

Особено значение се отдава на семантичното обогатяване на съдържанието чрез извличане на именувани същности, което позволява идентифициране на по-дълбоки връзки между публикациите, невидими при традиционните лексикални представяния, и допринася за повишаване на качеството и обяснимостта на препоръките. Услугата за **извличане на именувани същности** обработва информационните ресурси на партии и генерира обогатени метаданни под формата на структурирани списъци от лица, организации, места и други същности. В метаданните се включват само същности, които надхвърлят конфигурируем праг на честота и достоверност, което редуцира шума и запазва единствено елементите, релевантни за последващата оценка на сходство. Тези метаданни участват пряко в изчисляването на близостта между документите и допринасят за повишаване на точността, устойчивостта при оскъдни данни и обяснимостта на резултатите.

Първият процес извършва семантично представяне на текстовете и конструира **матрица на сходство** между документите, използвана както за модула за „подобни текстове“, така и като ядро на персонализиращия алгоритъм. Вторият процес анализира регистрите на взаимодействията със системата (напр. преглеждания и изтегляния), включително анонимни, и изгражда разредена **матрица „потребител-документ“**, както и **вектор на популярност на документите**. Данните се приемат поточно, а изчисленията се изпълняват периодично в пакетен режим, като при всяка итерация се обработват само новопостъпилите записи.

Обработката е организирана инкрементално, което позволява поетапно управление на големи масиви от данни и последващо обновяване единствено на новата информация. Например при добавяне на нови текстови ресурси не се изисква пълно преизчисляване на сходствата, а само оценка на отношенията спрямо новите елементи. Архитектурата и предварително изчислените структури са проектирани така, че да осигуряват устойчивост и съгласуваност на резултатите дори при включване на тематично и съдържателно различни данни.

Предлаганото решение осигурява унифицирана логика за представяне и обработка на данните чрез единни идентификатори, съгласувани процедури за обновяване и ниска латентност на отговорите. Архитектурата включва ясно разграничени компоненти - отделна услуга за извличане на именувани същности и два модула за персонализирано представяне на съдържание: „подобни документи“ и „персонализирани препоръки“.

Интерактивната част на системата се реализира чрез два модула, които стъпват върху предварително изчислените оперативни структури, осигурявайки ниска латентност, последователност и възпроизводимост на резултатите. Първият модул визуализира **„подобни текстове“**, като при достъп до конкретен ресурс се извличат топ-*k* най-сходни документи въз основа на съответния ред от матрицата на сходство и зададен праг. Този механизъм е инвариантен спрямо потребителя и осигурява стабилна и обяснима ориентация в корпуса. Вторият модул генерира **персонализирани препоръки**, използвайки индивидуалната история на взаимодействията, когато тя е налична. От разредената матрица „потребител-документ“ се формира вектор на предпочитанията, който се прилага върху матрицата на сходство за претегляне на кандидатите, като резултатът се комбинира с показател за популярност. По този начин препоръките отчитат както съдържателната близост до вече разглеждани ресурси, така и

глобалната им значимост. В условия на студен старт или при липса на релевантни взаимодействия системата временно преминава към препоръчване на най-представителните ресурси по популярност до натрупване на достатъчно данни.

Изградената архитектура е насочена към целенасочено смекчаване на ограниченията, идентифицирани в литературния преглед, включително зависимостта от богати поведенчески данни, проблема със студения старт, мащабируемостта на изчислително тежки модели, ограничената обяснимост и фрагментарното използване на наличните източници на информация. Изследователската рамка акцентира върху практическата приложимост в големи и разрастващи се колекции, като изчислително интензивните стъпки се изнасят в етапа на подготовка на данните, а интерактивният слой използва компактни, предварително изчислени структури с инкрементално обновяване.

В следващите подраздели се разглеждат в детайли отделните модули и услуги, използваните данни и тяхната подготовка, както и алгоритмичните стъпки и механизмите за периодично обновяване.

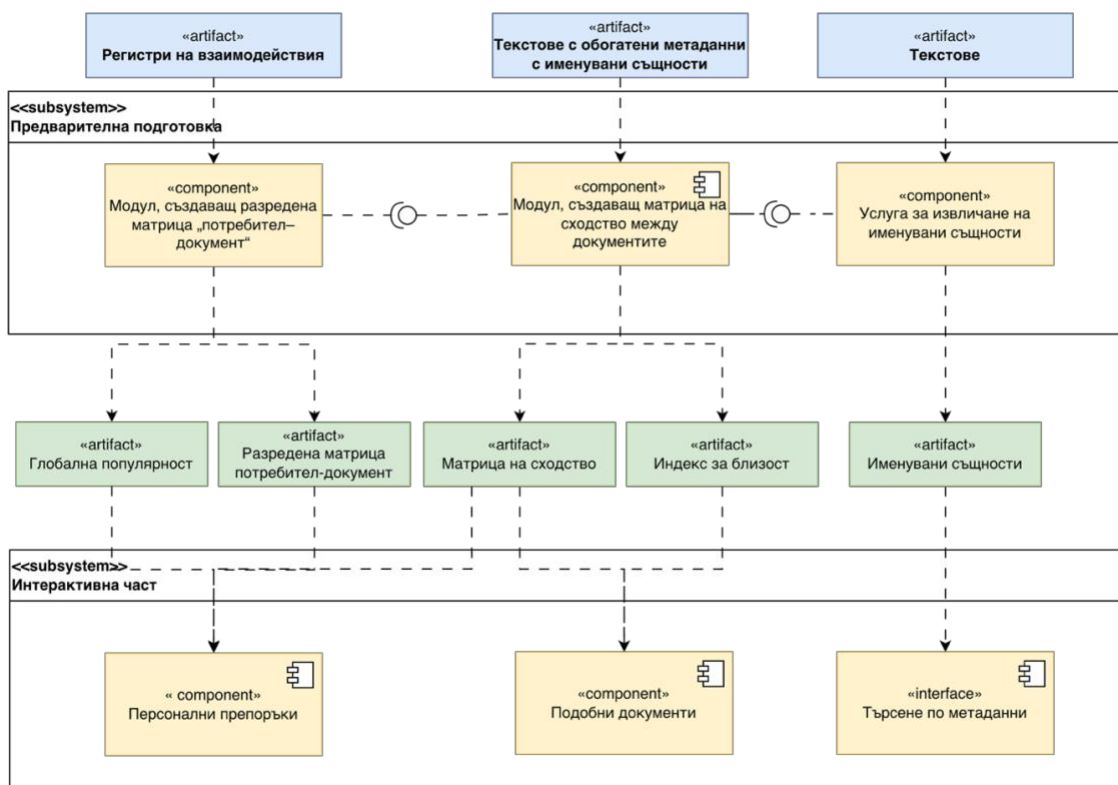
3.2. Концептуален модел и архитектурна рамка за персонализирано представяне на съдържание в дигитална библиотека

3.2.1. Концептуален модел

Концептуалният модел за персонализирано представяне на съдържание в дигитална библиотека дефинира основните компоненти на системата и техните роли в цялостната архитектура. Моделът разграничава етап на **предварителна подготовка** (вж. фиг. 9), който функционира извън непосредственото взаимодействие с потребителя и подготвя необходимите структури и данни, и **интерактивен етап**, който използва тези предварително изчислени структури за реализиране на хибридният алгоритъм за препоръчване с кратко време за отговор (вж. фиг. 9).

Концептуалният модел се основава на два взаимодопълващи се източника на данни (показатели) - **текстовите ресурси** и **регистрите на взаимодействията на потребителите със системата** (вж. фиг. 9). От текстовите ресурси се извличат **именувани същности**, които формират допълнителен семантичен слой и се използват като трети източник на данни. Въз основа на тези данни се изграждат компактни и съгласувани оперативни структури, подлежащи на инкрементално обновяване при постъпване на нови ресурси или нови записи в регистрите за взаимодействие. Тази

организация осигурява възпроизводимост на резултатите, мащабируемост и устойчиво персонализиране в динамична среда.



Фигура 9. Компонентна диаграма на концептуалния модел

Понятиятният слой включва обектите „документ“, „потребител“, „събитие за достъп“ и „именувана същност“.

- „Документ“ представлява текстов ресурс, съхранен в дигиталната библиотека, като в текущата реализация се разглеждат единствено периодични издания на български език.
- „Потребител“ се представя чрез анонимен идентификатор, съобразен с изискванията за защита на личните данни.
- „Събитие за достъп“ описва взаимодействието между потребител и документ, допълнено с времеви печат и тип действие (напр. преглед, изтегляне).
- „Именувана същност“ е структурирано представяне на обект от текста (лице, организация, място и др.), което обогатява метаданните и при необходимост участва в оценката на близостта между документите.

Връзките между обектите се реализират чрез единни идентификатори, което гарантира съгласуваност между отделните модули на системата.

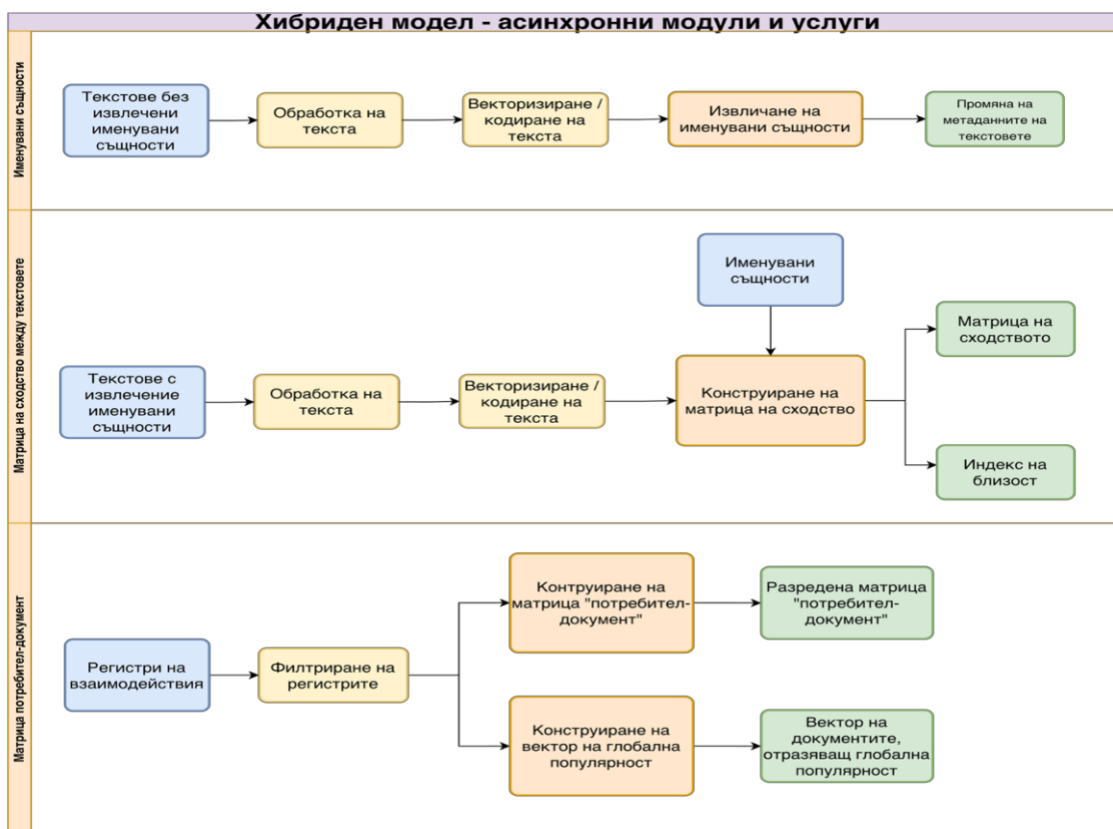
От трите вида показатели - съдържателни, поведенчески и семантични - се изграждат основните оперативни структури на системата:

1. матрица на сходство между документите;
2. разредена матрица „потребител-документ“;
3. индикатор за глобална популярност на документите.

Тези структури формират основата за навигация по смислова близост и за персонализирано предлагане на съдържание.

На фигура 10 са представени услугите и модулите, реализирани в етапа на предварителна подготовка. Показани са трите основни изчислително интензивни операции, изпълнявани асинхронно:

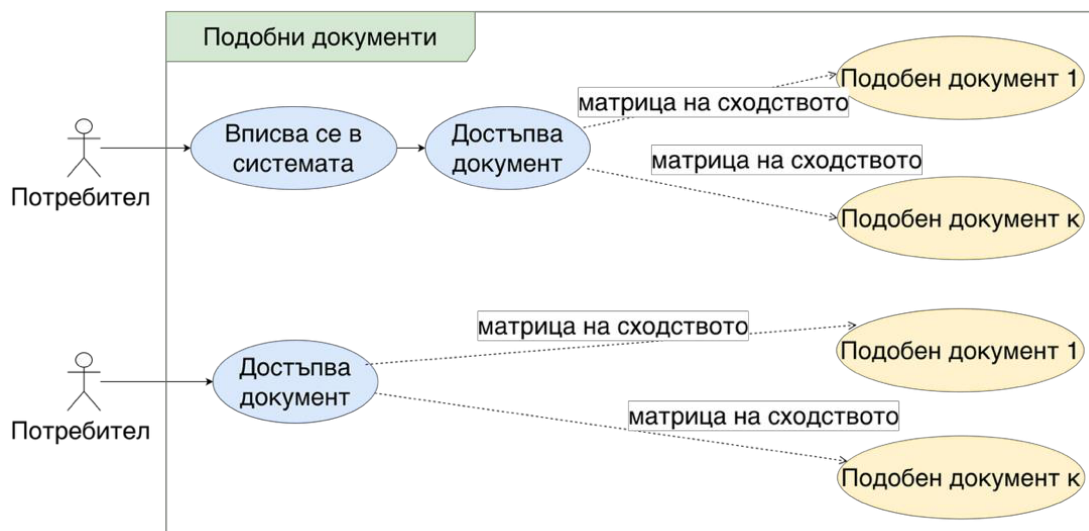
- Извличане на именувани същности от документите, реализирано в партиден режим;
- Изчисляване на матрица на сходство между документите чрез преобразуване на текстовете в числови представяния и съхраняване на резултатите в разреден вид като списъци с K най-близки съседи;
- Изграждане на разредена матрица „потребител-документ“ от регистрите на взаимодействията, при което честотата на достъп се интерпретира като имплицитна тежест и паралелно се изчислява показател за глобална популярност.



Фигура 10. Асинхронен слой

Етапът на предварителна подготовка генерира предварително изчислени структури и индекси, които се съхраняват и се използват от интерактивния слой. Интерактивният слой не изпълнява изчислително тежки операции, а работи изцяло върху готовите представяния, което осигурява кратко време за отговор и стабилност на резултатите. В рамките на този слой се реализират два допълващи се режима: **„подобни документи“** и **„персонализирани препоръки“**.

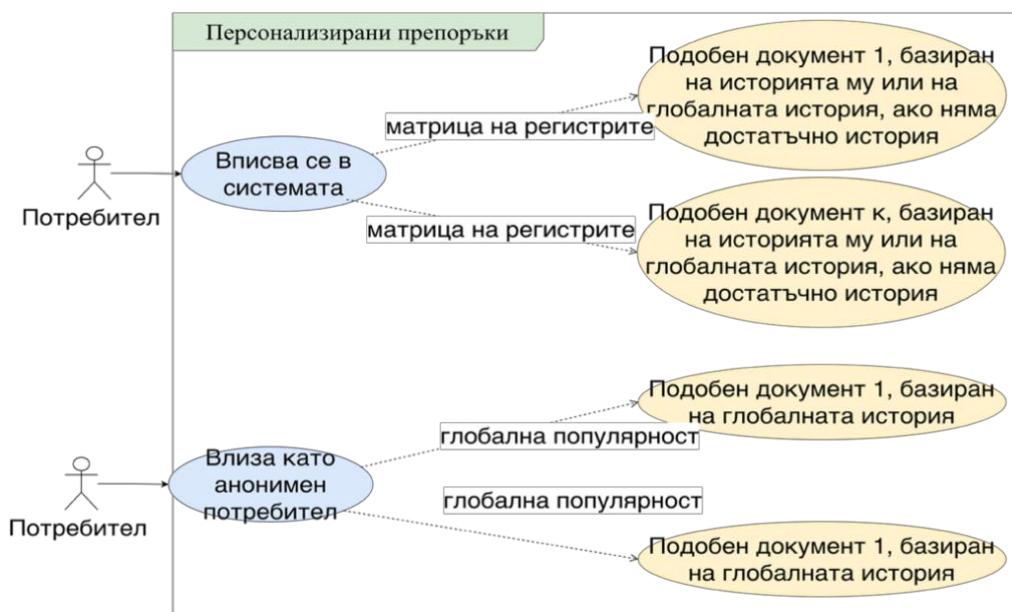
В режима **„подобни документи“** системата локализира активния ресурс по идентификатор, извлича съответния ред от матрицата на сходство и от индекс на най-близките съседи, прилага контролни прагове и филтри и връща документите с най-високи оценки за смислова близост (вж. фиг. 11).



Фигура 11. Диаграма на случаи на употреба за „подобни документи“

В режима *„персонализирани препоръки“* интерактивният слой агрегира информацията от историята на потребителя, проектира предпочитанията към сходни по съдържание документи чрез матрицата на сходство и прилага коректив от индикатора за популярност. При условия на студен старт се използва комбинация от съдържателна близост и представителни по популярност ресурси (вж. фиг. 12). Крайната оценка се формира като разлагаема сума от компонентите, което улеснява обяснимостта на резултатите.

И в двата режима интерактивният слой оперира единствено върху предварително изчислените структури, като по този начин се *гарантират ниска латентност, мащабируемост и прозрачни основания за всяко предложено съдържание.*



Фигура 12. Диаграма на случаи на употреба за „Персонализирани препоръки“

3.2.2. Архитектурна рамка

Както вече бе отбелязано, предложеното решение се основава на многослойна архитектура, при която *изчислително интензивните операции се изпълняват предварително*, а при обслужване на потребителите системата използва *предварително изчислени представяния, индекси и агрегирани показатели*. Тази организация е насочена към три свързани цели: (1) *да се осигури мащабируемост, като сложните и тежки изчисления се изнесат извън критичния път на потребителските заявки*; (2) *да се намали зависимостта от богата индивидуална история чрез силен съдържателен компонент и така да се смекчи проблемът на „студения старт“*; и (3) *да се постигне по-висока обяснимост чрез разделими и проследими компоненти (сходство по съдържание, поведение, популярност)*. В този контекст архитектурата интегрира съдържателни, поведенчески и структурни данни в съгласуван модел.

Архитектурата включва следните основни нива.

1. Ниво „Съдържание“ - на това ниво е **подбран текстов корпус на български език** от периодични издания, съхранени в структурирани записи.
 - а. От всяка структура се извличат самостоятелни текстови единици, като за всяка се използва идентификатор.

- б. За всеки документ се прилага единна процедура за предварителна обработка (подготовка на текстовите данни).
 - в. Дългите текстове се разделят на по-къси фрагменти, за да се улови вътрешното тематично разнообразие.
 - г. Всеки фрагмент се представя чрез числов вектор, изчислен от обучен езиков модел. Така всеки документ получава компактно многомерно представяне, отразяващо неговото съдържание.
 - д. На това ниво текстовете се „превеждат“ във векторна форма, подходяща за измерване на сходство и за последващи изчисления.
2. Ниво „Матрица на сходство между документите“ - *на базата на векторните представяния се изгражда матрица на сходство между всички документи.*
- а. Комбинират се различни аспекти на близост:
 - 1) сходство между обобщените представяния на документите;
 - 2) максимално сходство между отделни фрагменти (за улавяне на локални тематични съвпадения);
 - 3) по избор - прилагане допълнително на метода за размитите к-средни, за да може да се определи частичната принадлежност на документите;
 - 4) по избор - допълнителен принос от съвпадащи именувани същности (лица, институции, географски названия и др.), извлечени автоматично и филтрирани според честотата им в корпуса.
 - б. Получената матрица на сходство е **симетрична** и обобщава както общата тематика, така и по-специфичните връзки между текстовете.
3. Ниво „Персонализирани препоръки“ - *тук се използват натрупаните данни от реално ползване на системата.*
- а. От регистрите за взаимодействия на потребителите със системата се извличат събития за достъп до ресурсите.
 - б. На тази основа се изгражда разредена матрица „потребител-документ“, която отразява кои документи са преглеждани от даден потребител и колко

пъти. Броят на прегледите се преобразува в плавна оценка (имплицитно одобрение - няма налични реални оценки).

- в. Паралелно се изчислява обща популярност на всеки документ, която отразява интереса на цялата аудитория, включително достъпът от анонимни потребители.
4. Ниво „Хибриден алгоритъм за препоръчване“ - *върху матрицата на сходство и матрицата „потребител-документ“ се изгражда хибриден алгоритъм за препоръчване*, който работи по следния начин:
- а. При потребители с **натрупана история** системата използва сходството между документите, за да открие текстове, **близки до вече достъпените**;
 - б. При **нови, слабо активни потребители** или потребители, които са изчерпали възможните препоръки на текущата им история се използват **най-популярните и представителни документи**, като по този начин се избягва „студен старт“ и се осигурява базово качество на препоръките.
 - в. Модулът „Подобни текстове“ използва **директно ред от матрицата на сходство** за конкретния документ и визуализира списък от най-близки текстове под него.
5. Ниво „Актуализация и мащабиране“ - архитектурата е проектирана така, че да *поддържа нарастващ обем от документи и потребителска активност*.
- а. Новите документи периодично се включват в корпуса чрез извличане, обработка и актуализиране на представянията и сходствата.
 - б. Новопостъпилите регистри на взаимодействия със системата се агрегират в актуализирана разредена матрица „потребител-документ“ и във вектор на глобална популярност, така че препоръките да отразяват динамиката на реалните интереси.
 - в. Тежките операции по векторизиране и изчисляване на сходство могат да се изпълняват по график и паралелно, докато интерактивната част остава лека.

Така архитектурната рамка проследява ясен поток: от сурови текстове и регистри на взаимодействия, през унифицирани числови представяния, матрица на сходство и

матрица „потребител-документ“, до услуги за „подобни текстове“ и персонализирани препоръки, с възможност за поетапно надграждане според нуждите.

В следващите части архитектурната рамка ще бъде разгърната в по-голяма детайлност.

3.3. 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 **обогатява 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 на 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 е проектирана като **самостоятелна, асинхронна услуга, която работи в пакетен режим и обработва информационните ресурси на партиди**. Извличането на именуваните същности става с използване на **голям езиков модел**, но преди неговото прилагане, текстовете преминават през минимална предварителна обработка. Обработката е минимална и не трябва да включва промяна на регистъра на буквите и премахване на пунктуация, *за да не се компрометира способността за коректно разпознаване на именувани същности*.

Получените списъци от именувани същности се съхраняват като част от обогатените метаданни на всеки документ и изпълняват три взаимосвързани функции:

- **подпомагат търсенето**, като позволяват заявки по конкретни лица, институции или места и предоставят по-прецизно подреждане на резултатите;
- служат като **допълнителен, лесно интерпретируем индикатор** в матрицата на сходство, където документи със значително припокриване на именувани същности получават по-висока оценка за близост;

- осигуряват **съдържателна основа** за препоръки.

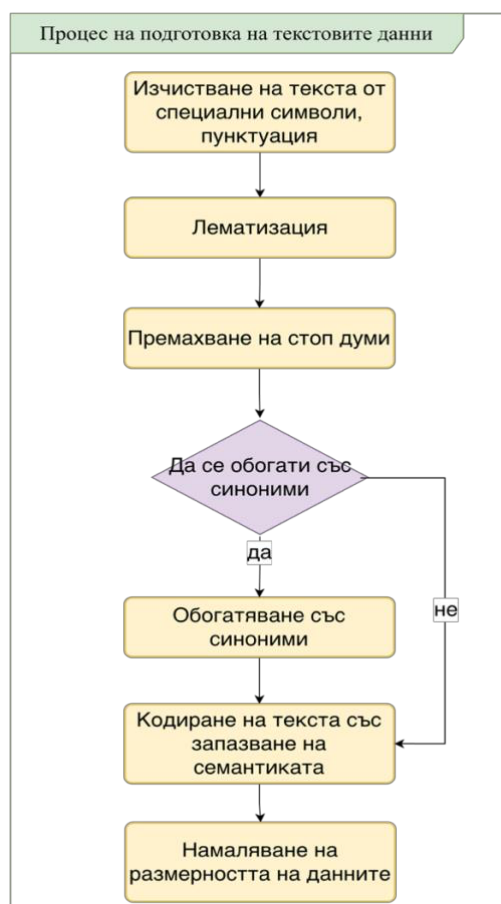
По този начин услугата за именувани същности се вписва в архитектурната рамка като хибриден мост между текстовото съдържание, метаданните и поведенческите данни.

3.4. 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**, преобразувани във векторни представяния след стандартизирана подготовка, и (2) извлечените от тях **именувани същности**. Върху това се дефинира оценка, която съчетава **косинусова близост** между векторите с **индикатор от съвпадения по именувани същности**; резултатът се организира като **компактна 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.

3.4.1. Подготовка на текстовите данни

Етапът на подготовка на текстовите данни (вж. фиг. 13) има за цел да осигури надеждна и консистентна основа за изчисляване на сходство. Най-напред се **редуцира шумът** и се отстраняват **нерелевантни елементи** (препинателни знаци, числа, специални символи), а текстът се нормализира чрез уеднаквяване на регистъра. След това се прилага **лематизация**, която свежда думите до базова форма, намалява разредеността на представянето и стабилизира мерките за близост между текстове [167], което е последвано от премахване на така наречените **стоп-думи**, които не носят релевантна информация, но увеличават и без това големия обем на документите. По избор се добавя **синонимно обогатяване**, за да се смекчат повърхностните лексикални вариации.



Фигура 13. Подготовка на текстовите обекти

На следващ етап корпусът се преобразува в числова форма, подходяща за по-нататъшни изчисления и прилагане на методи от изкуствения интелект и машинното обучение.

3.4.2. Векторизация на текстови данни

Целта на векторизацията е да трансформира подготвените текстове в числови представяния, подходящи за алгоритмите от изкуствения интелект и машинното обучение. Входът към този етап са нормализирани, лематизирани и синонимно обогатени текстове (по избор); изходът е еднороден набор от вектори.

Класическите векторизации като „чанта от думи“ и **TF-IDF** [168] предлагат прозрачно, но повърхностно кодиране, тъй като не отчитат семантичната близост между думите. Това ограничение се преодолява чрез **вграждания** (embeddings) - векторни представяния, които кодират *семантичната и контекстуалната близост между думите* [98], [99]. При тях всяка дума се представя като вектор в многомерно

пространство, така че лексеми с подобно значение или контекст да бъдат разположени близо една до друга. Статични модели назначават еднакъв вектор на думата независимо от контекста [103], [104]. **Контекстуалните вграждания** преодоляват тези ограничения, като позиционират семантично сходни изрази близо един до друг в общо пространство [105] и така повишават точността на последващите операции по сходство, групиране и препоръчване, което повишава семантичната изразителност и обобщаващата способност [102].

За векторизация на данните е използван подход, базиран на **контекстуални вграждания** чрез модели от архитектура „трансформър“, които кодират както лексикална, така и контекстно-семантична информация. Приложен е компактен модел с **384-измерни вектори**, поддържащ български език и осигуряващ ефективен баланс между качество и изчислителна производителност. Ограниченията в максималната дължина на входната последователност се преодоляват чрез *сегментиране на документите в застъпващи се фрагменти и последваща агрегация на ниво документ, което запазва глобалната тематична структура*. Допълнително **намаляване на размерността не се прилага**, тъй като то води до нестабилност и несъпоставимост при инкрементална обработка, докато директното използване на плътните вграждания гарантира стабилност, консистентност и възпроизводимост на представянията.

3.4.3. Намаляване на размерността при текстови данни

Прилагането на методи от машинното обучение и изкуствения интелект за изграждане на персонализирани препоръчителни системи в дигитални библиотеки съвсем не е тривиална задача поради големия обем и тематичното разнообразие на текстовите данни, с които тези системи оперират. Документите в подобни среди обикновено са дълги, многотематични и с висок степен на семантично припокриване, което прави коректното измерване на сходство трудно, а изчислителната цена нараства съществено.

За преодоляване на тези предизвикателства се използват техники за **намаляване на размерността на текстовите данни**, чиято цел е редуциране на броя признаци без загуба на съществена информация за семантичната структура [98]. Така се получава по-компактно и устойчиво представяне на текстовите вектори, елиминират се шум и излишни корелации, намалява се рискът от пренапасване и се улесняват визуализацията и интерпретацията на резултатите [169]. Съвременните изследвания [170], [171]

показват, че при дълги и тематично разнообразни текстове, характерни за дигиталните библиотеки, ефективни са както линейни методи като PCA и NMF, така и нелинейни подходи като UMAP, които улавят локалната и глобалната структура на данните.

В този контекст NMF подпомага извличането на латентни теми, UMAP се използва за клъстеризация и визуализация [172], а PCA - за предварителна редукция на признаците. Допълнителни методи като t-SNE и автоенкодери намират приложение при визуализация и нелинейно компресиране на векторни представяния, като подпомагат тематичното групиране и анализа на данните [173], [174], [175], [176].

В таблица 8 е направен сравнителен анализ на предимствата и недостатъците на различните методи за намаляване на размерността на текстовите данни [170], [171], [172], [173], [175], [176].

Таблица 8. Сравнение на техники за намаляване на размерността при дълги текстове

<i>Метод и Тип</i>	<i>Приложимост при текст</i>	<i>Предимства</i>	<i>Ограничения</i>	<i>Препоръчителна употреба</i>
<i>Анализ на главните компоненти (PCA) -- Линеен</i>	Подходящ при представяния чрез претеглено представяне или вграждания.	Намалява шум и корелации между признаците; осигурява по- интерпретируемо признаково пространство; бърз и ефективен при големи корпуси.	Не улавя нелинейни зависимости; може да загуби локална структура на данните.	Използва се при големи корпуси с ясно изразени теми и за предварителна намаляване на пространството преди обучение на модели.
<i>Факторизация на неотрицателни матрици (NMF) --</i>	Особено подходящ за извличане на теми в текстови данни.	Интерпретируеми компоненти; често използван при тематично моделиране; стабилен при	Изисква неотрицателни стойности; не гарантира глобален минимум.	Използва се за анализ на теми, клъстеризация на документи и определяне на теми при

<i>Метод и Тип</i>	<i>Приложимост при текст</i>	<i>Предимства</i>	<i>Ограничения</i>	<i>Препоръчителна употреба</i>
<i>Линеен / тематичен</i>		претеглено представяне.		периодични издания.
<i>t-разпределено стохастично вграждане на съседи (t-SNE) -- Нелинеен</i>	Използва се за визуализация на документи в 2D/3D пространства.	Отлично запазва локалните отношения и сходства между документи;	Висока изчислителна сложност; трудно се скалира за големи корпуси; чувствителен към параметри.	Подходящ за визуализация и изследване на тематични кълъстери и връзки между статии.
<i>Равномерна апроксимация и проекция на многообразието (UMAP) -- Нелинеен</i>	Добър баланс между скорост и качество при текстови вграждания.	По-бърз и стабилен от t-SNE; запазва както локалната, така и глобалната структура; добре се скалира.	Чувствителен към избор на параметри (напр. брой съседи, минимално разстояние).	Подходящ за кълъстеризация и визуализация на големи корпуси с разнообразна тематика.
<i>Автоенкодер -- Дълбок / невронен</i>	Приложим при компресиране на вграждания.	Улавя сложни нелинейни зависимости; адаптивен към различни типове текстови представяния; възможност за дообучаване.	Изисква големи изчислителни ресурси; трудно интерпретируем; риск от пренапасване.	Използва се за семантична кълъстеризация, извличане на латентни теми и компресия на високоразмерни вграждания.

Въпреки безспорните теоретични предимства на изброените методи, при проектирането на архитектурата за конкретната дигитална библиотека бе извършен критичен анализ на тяхната приложимост спрямо изискването за инкрементално обновяване. Методите за редукция (като PCA и LSA) изграждат проекционни матрици,

зависими от статистическото разпределение на първоначалния обучителен корпус. При динамично постъпване на нови документи от непознати до момента тематични области (domain drift), тези статични модели губят своята представителност и налагат пълно преизчисляване на индекса, което е неефективно за реални дигитални библиотеки и предложената архитектура.

Поради тази причина, в настоящата дисертация е възприет алтернативен подход. Както беше споменато в предната подглава вместо математическа редукция на размерността, която носи риск от загуба на семантичен детайл при нови данни, се използва модел за кодиране на данните, който намалява и размерността им. Избраният модел (MiniLM) генерира сравнително компактни вектори в сравнение с класическите разредени матрици. Този подход осигурява намаляване на необходимата памет с 50%, като същевременно запазва напълно семантичната разделителна способност и позволява линейно добавяне на нови документи без необходимост от преизчисляване на индексите.

3.4.4. Матрица на сходство и вход към „подобни текстове“

В тази подглава се представя в детайли **матрицата на сходство между документите** като формализирано представяне на **смисловата близост** в корпуса. Матрицата се конструира на базата на метода на **многокомпонентната оценка за близост** (семантични векторни представяния и съвпадение по именувани същности) и се запазва като **разреден индекс от най-близки съседи (top-k)**, което ограничава паметта и ускорява достъпа.

Семантично представяне на документите и компоненти на сходство

Всеки документ i се представя като последователност от фрагменти, получени след предварителна нормализация и лематизация на текста. За всеки фрагмент се изчислява векторно представяне чрез предварително обучен езиков модел за български език. По този начин документът се описва като множество от семантични вектори $\{c_{i,k}\}$, което позволява улавяне както на глобалната тематика, така и на локалните подтеми.

Върху тези представяния се дефинира набор от *допълващи се компоненти за оценка на сходството* между документи i и j , целящи да обхванат различни аспекти на смисловата близост. **Глобалният компонент** осигурява *обобщена и устойчива мярка за тематично сходство*, докато **локалният** (по фрагменти) открива *силни тематични съвпадения в рамките на дълги или многотематични текстове*. **Тематичният компонент** въвежда *по-груба, но интерпретируема структура чрез групиране в широки*

тематични области, а сходството по **именувани същности** добавя *структуриран семантичен сигнал*, особено информативен при периодични издания.

Комбинирани параметризирано, тези компоненти формират балансирана, точна и обяснима оценка на сходството, като всеки допринася различна перспектива към смисловата близост. В следващото изложение всеки компонент е формализиран чрез конкретна метрика и включен в общата функция за сходство.

1. **Глобално (усреднено) сходство** между документите се изчислява като косинусова сходност между техните усреднени векторни представления. Повисоката стойност на мярката индикира по-голяма семантична близост и по-висока степен на сходство по съдържание:

$$S_{\text{усреднено}}(i, j) = \cos(\bar{c}_i, \bar{c}_j) = \frac{\bar{c}_i * \bar{c}_j}{\|\bar{c}_i\| * \|\bar{c}_j\|}$$

2. **Локално сходство** - 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 разграничими тематични блокове, дори когато глобалната им тематика се разминава.

$$S_{\text{най-добро}}(i, j) = \max_{k \in \{1, 2, \dots, n_i\}, l \in \{1, 2, \dots, n_j\}} \cos(c_{i,k}, c_{j,l})$$

3. **Тематична близост** на база „размито“ клъстеризиране (по избор) - всички фрагментни вектори се групират чрез алгоритъм за размито клъстеризиране (fuzzy c-means) в C тематични центъра. Всеки фрагмент получава степен на принадлежност $u_{t,c} \in [0, 1]$ към темите. Тематичният профил на документ i е усреднен вектор на принадлежностите на неговите фрагменти:

$$t_i = \frac{1}{n_i} \sum_{k=1}^{n_i} (u_{(i,k),1}, \dots, u_{(i,k),C})$$

Тематичното сходство се измерва отново с косинусова мярка, като тази компонента е активна само ако тематичният компонент е включен.

$$S_{\text{тематично}}(i, j) = \cos(t_i, t_j) = \frac{t_i * t_j}{\|t_i\| * \|t_j\|}$$

4. **Сходство по именувани същности** - 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** - тази компонента дава допълнителна близост на документи, които споделят съществени общи обекти.

$$S_{\text{именувани същности}}(i, j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|}$$

Матрица на сходство

Съдържателната оценка се формира на базата на глобално, локално и тематично сходство, които се обединяват в следната многокомпонентна мярка:

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

Където $\alpha, \beta, \gamma > 0$ и $\alpha + \beta + \gamma = 1$, ако е включена тематичната близост на база „размито“ клъстеризиране, и съответно $\gamma = 0$, когато е изключено.

Когато се използва и информация от именувани същности, крайното сходство се задава като:

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

При $\lambda = 0$ се работи само със съдържателна близост; при ненулева стойност на λ именуваните същности, съхранени в метаданните, имат ролята на коригиращ и обясним фактор.

За да се ограничи паметта и да се ускори последващият достъп, вместо пълна матрица се съхранява разреден **к-най-близки съседни (k-NN) индекс**, при който за всеки документ се пазят само връзките с най-висока сходност (K на брой или над праг) [59]. Получената матрица $S_{\text{финална}}$ е **симетрична**, с размер $N \times N$, където N е броят на обработените до момента документите.

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

За да може тази съдържателна перспектива да бъде допълнена с реалното поведение на потребителите, е необходимо паралелно да се конструира и надежден модел на взаимодействията на потребителите със системата. В следващия подраздел се

разглежда именно подготовката на регистрите на взаимодействия със системата, чрез която суровите регистри се трансформират в структурирани имплицитни оценки, съвместими с вече дефинираната представителна схема за документите.

3.5. Матрица „потребител-документ“ и имплицитни оценки

Регистрите на взаимодействия между потребителите и документите, генерирани от 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 алгоритъм за генериране на персонализирани препоръки.

3.5.1. Филтриране и нормализиране на събитията

Първата стъпка при обработката на регистрите на взаимодействия на потребителите със системата е предварително филтриране на нерелевантните записи. От регистрите се изключват административни операции (създаване, промяна и изтриване на обекти), системни събития, които не носят информация за потребителското поведение.

След това релевантните записи се нормализират чрез редуциране до минимален набор от атрибути: идентификатор на потребителя (u), идентификатор на документа (i), времеви печат и тип действие. При конструирането на разредената матрица „потребител-документ“ се използват единствено събитията, отразяващи **достъп до ресурс**, тъй като те предоставят пряка информация за интересите на потребителите.

Останалите типове операции, като създаване, модификация или изтриване на ресурси, не участват пряко в матрицата, а се използват като сигнали за инкрементална актуализация на оперативните структури, включително матрицата на сходство и свързаните с нея индекси и речници.

3.5.2. 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 тегло $w_{u,i}$ чрез монотонно нарастваща, но насищаща се функция:

$$w_{u,i} = \begin{cases} 0, & c_{u,i} \leq 0 \\ \min(1 + 0.3 * (c_{u,i} - 1), 2), & c_{u,i} > 0 \end{cases}$$

Така единичното преглеждане задава базова тежест; допълнителните преглеждания я повишават с намаляващ прираст. На основата на тези тегла се конструира разредена матрица на взаимодействията. Матрицата се съхранява в уплътнен формат за разредени данни.

3.5.3. Вектор на глобална популярност на документите

Успоредно с индивидуалните взаимодействия се изчислява вектор на глобална популярност на документите, базиран на броя преглеждания, трансформирани чрез монотонно нарастваща, насищаща се функция и нормализиран в интервала [0,1]. Този показател се използва като допълнителен фактор в препоръчителния модел 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:

$$\text{популярност}(i) = \sum_u w_{u,i}$$

В следващия подраздел се разглежда съвместното използване на семантичния и поведенческият поток чрез изграждане на общи оперативни структури и индекси.

3.6. 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 структурни представяния (напр. индекси и матрици), надеждното им съхранение и процедури за периодично/инкрементално обновяване, така че системата да запазва коректност, устойчивост и производителност при нарастващ обем данни и натоварване.

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 представления.

В резултат на предварителната обработка се поддържат няколко основни оперативни структури, изградени извън линията на директното взаимодействие и обновявани периодично. Сред тях са **матрица на сходство** между документите, изчислена върху векторните представления и интегрираща глобална и локална семантична близост, както и допълнителни тематични и структурни показатели; **разредена матрица „потребител-документ“** и **вектор на глобална популярност**, които отразяват реалното използване на ресурсите; както и обогатени метаданни и извлечени **именувани същности**, използвани като допълнителен семантичен слой.

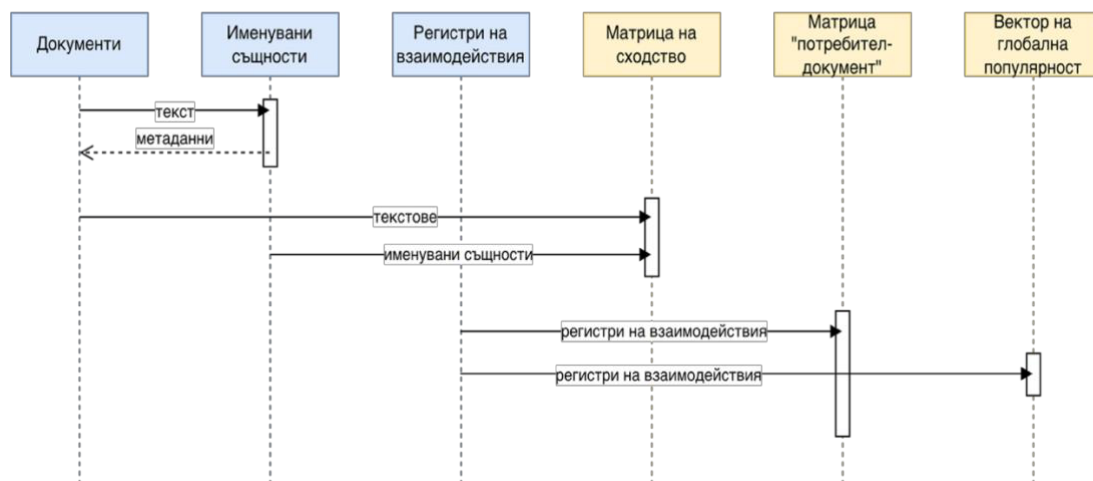
Системата регистрира всички типове взаимодействия, като за изграждането на *матрицата „потребител-документ“* се използват единствено събитията за достъп до ресурси, интерпретирани като имплицитен индикатор за интерес. Останалите събития не участват пряко в изчисленията, но се използват за **поддържане на актуалността и консистентността на представленията**. 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. 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.

3.7. Модули за генериране на персонализирано съдържание

В тази глава се представят два функционални модула за персонализирано препоръчване в дигитални библиотеки, които обединяват съдържателна семантична близост, наблюдавано потребителско поведение и глобална популярност на документите в единна, параметризируема рамка. Модулите целят да осигурят релевантни и обясними препоръки дори и при ограничена индивидуална история, като същевременно поддържат оперативна ефективност и мащабируемост при реални натоварвания.

За яснота на зависимостите и реда на изчисленията, диаграмата на последователността (вж. фиг. 14) показва как входните потоци - пълните текстове и регистрите на взаимодействия - се трансформират в предварително изчислени структури.



Фигура 14. Диаграма на последователността на създаване на оперативните структури

Последователността допуска успоредно изпълнение, но методически е целесъобразно първо да се извлекат именувани същности чрез отделната услуга, за да може последващата матрица на сходство документ-документ да включи и този допълнителен семантичен принос. Паралелно регистрите за взаимодействия със системата се преобразуват в разрежена матрица „потребител-документ“ и във вектор на глобална популярност.

Върху изчислените оперативни структури се реализират два функционални модула: модул за препоръчване на „подобни документи“, базирано единствено на съдържателната близост, и модул за генериране на персонализирани препоръки, който комбинира семантична близост, индивидуална история и глобална популярност. В следващите подглави последователно се конкретизират тези модули.

3.7.1. Функционален модул за „подобни документи“ и метод на многокомпонентната оценка

Функционалният модул за генериране на „подобни документи“ представя документ-центрирана услуга за навигация в корпуса, чиято цел е да извежда най-близките по съдържание информационни ресурси спрямо текущо разглеждания ресурс. Подходът е **потребителско-инвариантен**: резултатите се определят единствено от

смиловата близост между самите документи и затова остават стабилни и лесни за интерпретация.

За да се реализира тази функционалност, корпусът се проектира в семантично пространство и върху него се конструира **симетрична матрица на сходство S** . Оценката $s(i, j)$ комбинира (1) глобална семантична близост (косинус между усреднените представяния), (2) локални съвпадения (максимална близост между фрагментни двойки), (3) по избор тематични профили от „размито“ групиране и (4) принос от именувани същности. Матрицата се съхранява в разреден вид като индекс на **k -те най-близки съседи за всеки документ**, съгласуван с общия речник на идентификаторите.

Извличането е просто и бързо: системата *локализира активния документ по идентификатор, извлича съответния ред/списък със съседи, изключва самия документ и прилага праг 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.

3.7.2. Функционален модул за персонализирани препоръки и хибриден алгоритъм

Функционалният модул за генериране на „персонализирани препоръки“ представя **хибриден потребителско-центриран алгоритъм, който интегрира съдържателна близост, наблюдавано поведение и глобална популярност в единна, параметризируема схема за подреждане** (вж. фиг. 15).

В слоя за обработка на поведенческите данни регистрите за преглеждания се филтрират и агрегират до разредена матрица „потребител-документ“ W , в която взаимодействията се представят чрез имплицитни тегла. Теглата се изчисляват чрез монотонна, насищаща се трансформация на валидираните преглеждания, описана в раздел 3.3.2. Успоредно с това се формира вектор на глобална популярност p за документите. По този начин се моделира реалното потребление в системата и се осигурява устойчива стратегия при липса или оскъдност на индивидуална потребителска история, например при нови или анонимни потребители.

За всеки потребител u и кандидат-документ d оценката за релевантност се формира като сума от два компонента: (1) съдържателен принос, пренесен от личната история на потребителя чрез матрицата на сходство S ; (2) стабилизиращ компонент на

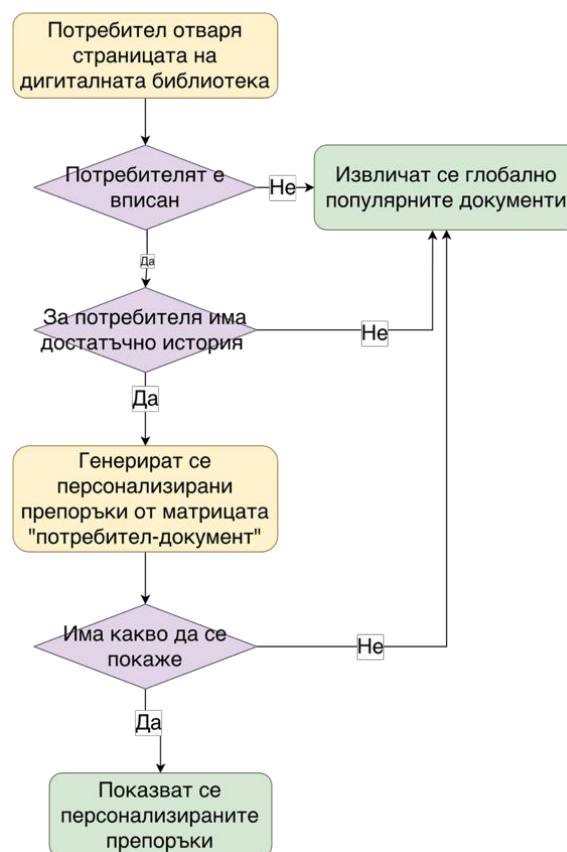
популярност $p(d)$, който се активира единствено при недостатъчна потребителска история. Като формално оценката се дефинира като:

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

където $\text{история}(u)$ е множеството документи, с които потребителят u е взаимодействал. Коефициентът $\partial(u)$ се определя адаптивно спрямо обема на наличната история:

$$\partial(u) = \begin{cases} \partial_0 \in [0.05, 0.20], & \sum_{i \in \text{история}(u)} w_{u,i} < H_{\min} \\ 0, & \text{иначе} \end{cases}$$

Популярността $p(d)$ се нормализира в интервала $[0,1]$ и функционира като резервна стратегия при студен старт, без да влияе върху оценката, когато личната история е достатъчно богата. В оперативен режим вече разглежданите от потребителя документи се изключват, а останалите кандидати се подреждат по $\text{оценка}(u, d)$.



Фигура 15. Концептуална схема на персонализирани препоръки

Алгоритъмът за генериране на персонализирани препоръки, работи в няколко режима. При потребители с **достатъчна история** документи, които са били прегледани,

формират ядро на профила, като за всеки кандидат d се оценява неговата близост до това ядро чрез матрицата на сходство. При **нови или анонимни потребители** препоръките се основават предимно на индикатора за популярност, като с натрупване на взаимодействия системата плавно преминава към пълната оценка без промяна на архитектурата. При **слабо активни потребители** хибридният алгоритъм осигурява стабилност, като съдържателният компонент остава водещ и компенсира оскъдните поведенчески връзки.

При **нови документи**, за които липсват натрупани взаимодействия („студен старт“ за елементи), съдържателният компонент предотвратява тяхната изолация, като ги включва в препоръките на потребители, чиито досегашни четения са семантично близки. Приносът на тези документи се регулира с по-ниско тегло до акумулиране на достатъчна поведенческа информация.

По този начин един и същ индекс на сходство служи като обща основа за двата ключови режима на работа: (1) извеждане на „подобни документи“; (2) пренасяне на личната история към нови, съдържателно сродни ресурси, допълнено при необходимост от индикатор за глобална популярност.

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.

3.8. Обяснимост и етични принципи при селекция на персонализирано съдържание

В съвременния контекст на персонализация в дигиталните библиотеки, където нараства използването на сложни алгоритми, въпросите за прозрачност, обяснимост и етично управление на данните придобиват все по-голяма значимост. Прегледът на изследванията показва, че значителна част от разработките поставят основен акцент върху точността на препоръките, докато аспекти като обяснение на използваните модели и по-последователното отчитане на етичните измерения остават по-слабо разгърнати [177], [178], [179]. Именно този установен пропуск обосновава включването на разглеждане на тези аспекти още при формулирането на архитектурните изисквания към персонализиращите решения в дигитални библиотеки.

3.8.1. Архитектурна обяснимост и интерпретация на резултатите

В предложената архитектура обяснимостта (Explainable AI) е присъща характеристика на хибридният модел (intrinsic explainability), заложена още на етапа на неговото проектиране. По този начин обосновката на препоръките произтича непосредствено от механизма за формиране на оценките, без необходимост от допълнителни слоеве за последваща интерпретация на резултатите. Това позволява да бъдат преодолені част от ограниченията, характерни за моделите тип „черна кутия“. Оценката за сходство се формира като разложима оценъчна функция, при която приносът на отделните компоненти е количествено обособим и смислово интерпретируем - глобално, локално и тематично сходство, както и фактологично припокриване чрез именувани същности. Тази архитектурна прозрачност осигурява проследимост на процеса на оценяване, предвидимо поведение при настройка на теглата и възможност за генериране на експлицитни обосновки за всеки резултат, например-например „предложено поради тематична близост“, „споделени именувани същности“ [177], [178], [180].

С цел избягване на погрешни интерпретации на суровите числови оценки от страна на потребителя, системата прилага слой за семантична трансформация, превеждайки стойностите на сходство в йерархични лингвистични етикети: „Идентичност“, „Високо сходство“, „Свързани“ и „Слаба връзка“. Границите между тези класове се калибрират емпирично върху валидираща извадка, като се анализира статистическото разпределение на оценките в матрицата и се верифицират примерни двойки документи. Тези прагове са дефинирани като конфигурируеми параметри и подлежат на периодичен преглед при промяна на данните или модела за векторизация, което гарантира устойчивост на етикетите във времето. Този дизайн е в пълно съзвучие с изследванията върху възприемането на обясненията и когнитивната достъпност на ХАІ подходите [181], [182].

В потребителския интерфейс се показва етикетът, а не абстрактното число; при необходимост интерфейсът позволява добавяне на кратко текстово основание. Този формат прави резултатите интуитивно сравними, избягва свръхинтерпретацията на пренебрежими числови разлики и улеснява разбирането защо даден документ е предложен. За научни и одитни цели пълните числови стойности се запазват в системните регистри, но не се изискват за ежедневната експлоатация на системата.

Подходът е приложим към разнородни колекции, като за всяка може да се поддържа собствен набор от прагове, съобразен с жанра и качеството на данните.

3.8.2. Минимизиране на алгоритмичните пристрастия и защита на данните

Проблемът с алгоритмичните пристрастия (algorithmic bias) се адресира на структурно ниво чрез разделяне на стратегиите за препоръчване. Влиянието на глобалната популярност е стриктно изолирано в ролята на резервна стратегия (fallback strategy) при липса на достатъчно история. Това архитектурно решение предотвратява усилването на ефекта на Матю („богатите стават по-богати“) [183] при персонализираните резултати, като не позволява на масовите документи да изместят специфичните за потребителя интереси.

Допълнително, рискът от манипулация на класацията чрез единични интензивни действия се минимизира чрез въведената функция на насищане при изчисляването на оценките в матрицата „потребител-документ“. Тази мярка ограничава тежестта на повторните интеракции и гарантира баланс между индивидуалните предпочитания и статистическия шум [179].

Паралелно с механизмите за обяснимост, системата прилага строги принципи за минимизация и целево използване на данните. Регистрите на взаимодействията съдържат единствено необходимия минимум от полета: псевдонимизиран потребителски идентификатор, идентификатор на документ, времеви печат и тип на действието [181], [184]. Извлечените именуванни същности преминават през филтрация по увереност и честота и се използват целево единствено за обогатяване на търсенето и обяснимостта, без да включват чувствителни категории. Достъпът до тях е ограничен по принципа „необходимост да се знае“ (need-to-know) и е проследим чрез диагностични регистри [184], [185]. Стриктното версионизиране на изчислените артефакти (матрици и индекси) осигурява проследимост и възпроизводимост на алгоритмичните решения [186].

3.9. Обобщение

В Глава 3 е разработена и аргументирана концептуална архитектура за персонализирано представяне на съдържание в дигитални библиотеки, която интегрира три допълващи се източника на знание: (1) съдържателна близост между документите, моделирана чрез семантични представяния, локални съвпадения, тематични профили и именуванни същности; (2) поведенчески показатели, извлечени от регистрите за

потребителски взаимодействия и формализирани чрез разредена матрица „потребител-документ“; и (3) метаданни и индикатори за глобална популярност.

Архитектурата изнася изчислително тежките операции в асинхронен слой, включващ предварителна обработка и нормализация на текстовете, векторизация, изграждане на матрица на сходство между документите, както и конструиране и на матрицата „потребител-документ“. Това позволява интерактивният слой да работи изцяло с предварително изчислени структури и да осигурява ниска латентност при генериране на резултатите.

Матрицата на сходство и матрицата „потребител-документ“ изпълняват съвместна роля в персонализиращия процес. Първата служи както за директно извеждане на „подобни документи“, така и като средство за пренасяне на предпочитанията, базирани на матрицата „потребител-документ“, към нови, съдържателно сродни ресурси. По този начин индивидуалната история на взаимодействията се проектира върху целия корпус от документи. При оскъдни или липсващи поведенчески данни този алгоритъм се допълва от индикатор за глобална популярност, който стабилизира подреждането на резултатите.

Описани са и механизмите за инкрементално обновяване и съгласуване на оперативните структури чрез единен индекс, което гарантира възпроизводимост, устойчивост и мащабируемост при нарастващ обем съдържание и потребителски взаимодействия.

Предложената архитектура е целенасочено проектирана да адресира ключови ограничения, идентифицирани в литературния преглед: зависимостта от оскъдни поведенчески данни (чрез комбиниране на матрицата „потребител-документ“ със съдържателни показатели и именувани същности), затруднената мащабируемост на сложни модели (чрез изнасяне на изчислително тежките операции в асинхронен слой и използване на предварително изчислени структури), недостатъчната обяснимост (чрез разложима многокомпонентна оценка с ясно проследими приноси) и фрагментарното използване на наличните източници на информация (чрез интегриране на съдържание, поведение и метаданни в рамките на единна архитектура).

В рамките на главата са формулирани и две изследователски хипотези, които се подлагат на емпирична проверка в следващата глава:

- **Хипотеза за многокомпонентна съдържателна оценка (матрица на сходство).**
Интегрирането на глобално семантично сходство, локални съвпадения, тематични профили и именувани същности води до по-точно и по-стабилно измерване на сходството между документи от многотематични периодични издания в сравнение с използването на отделни компоненти самостоятелно.
- **Хипотеза за хибридният алгоритъм при персонализирани препоръки.**
Комбинирането на съдържателната близост, матрицата „потребител-документ“ и адаптивен индикатор за глобална популярност води до по-качествени персонализирани препоръки, особено при студен старт, оскъдна история и тематични преходи, спрямо подходи, основани на един-единствен тип показатели.

ГЛАВА 4. ЕКСПЕРИМЕНТАЛНО ВНЕДРЯВАНЕ И АНАЛИЗ НА РЕЗУЛТАТНОТО ТЕСТВАНЕ

В тази глава се представя реализацията на модули за персонализирано представяне на съдържание в дигитална библиотека, в съответствие с концептуалния модел и архитектурата, предложена в глава 3. Персонализацията се основава на два основни показателя: (1) съдържателна близост между документите и (2) наблюдавано поведение на потребителите. Към тях се добавя и произведен семантичен показател от именувани същности, получен чрез отделна услуга за извличане на същности.

Реализацията е структурирана около два основни функционални модула и една самостоятелна услуга: (1) модул за конструиране на матрица на сходство, който обслужва функционалността „подобни документи“; (2) модул за обработка на регистрите на взаимодействия и извеждане на вектор на глобална популярност за „персонализирани препоръки“; и (3) услуга за извличане на именувани същности с цел обогатяване на метаданните и използването като допълнителен показател при конструиране на матрицата на сходство.

В следващите подраздели са описани входните и изходните структури, алгоритмичните стъпки и механизмите за интеграция между модулите, процедурите за калибриране на параметрите и верифицирането на модулите. Всеки подраздел, описващ един компонент от архитектурната рамка, завършва с експериментални резултати върху синтетични и реални данни, които демонстрират практическата приложимост на реализацията и служат за емпирична проверка на формулираните в предходната глава хипотези.

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

Програмната реализация стъпва върху модулна архитектура, съобразена с изискванията за високопроизводителни матрични изчисления и специфична езикова поддръжка за български език. Изборът на технологичен стек е мотивиран от необходимостта за ефективна работа с големи корпуси, възпроизводимост на резултатите и лесна интеграция между асинхронните процеси за предварителна подготовка и леката интерактивна част.

- Езици за програмиране и основна среда - Основен език, който се използва за разработката, е Python 3.13+, поради доминиращата му роля в екосистемата на машинното обучение, богатия набор от специализирани библиотеки и удобната интеграция между отделните слоеве на системата.
- Библиотеки за изкуствен интелект и машинно обучение - Използват се за реализацията на основните алгоритмични компоненти са използвани следните специализирани библиотеки:
 - sentence-transformers [187] (базирана на Hugging Face Transformers): Използвана за зареждане и управление на дълбоките невронни мрежи (моделът MiniLM), както и за генериране на семантичните векторни представяния.
 - scikit-learn [188]: Предоставя инструменти за предварителна обработка на данните и метрики за косинусово сходство.
 - scikit-fuzzy [189]: Специализирана библиотека за размита логика, използвана за имплементацията на алгоритъма на размитите к-средни (Fuzzy C-Means) при тематичното моделиране.
 - simplemma [190]: Лек и бърз лематизатор с поддръжка на български език, използван за нормализация на текстовия корпус и повишаващ точността на лексикалните сравнения.
 - stopwords-bg [191]: Набор от български стоп-думи за филтриране на неинформативни термини и подобряване на качеството на текстовата обработка.
- Инфраструктура за математически изчисления и оптимизация - Използва се за осигуряване на скалируемост и ефективност при работа с големи масиви от данни, изчислителното ядро на системата се базира на:
 - NumPy [192] и SciPy [193]: за линейна алгебра и разреждени структури; използва се Compressed Sparse Row (CSR) за ефективно съхранение и пресмятане, включително при коефициента на Жакар върху множествата от именувани същности;
 - PyTorch [194]: Служи като фундамент за тензорните изчисления на Transformer моделите.

- Хардуерна конфигурация и ускорение - Реализацията поддържа MPS (Apple Silicon) и CUDA (NVIDIA) за ускорение. За намаляване на паметния отпечатък и дисковия обем семантичните вектори и междинни представяния се съхраняват в полупрецизен формат (Float16), което редуцира размера с около 50% при пренебрежима загуба на точност в оценките на косинусова близост.
- Корпус и тестови данни:
 - За емпиричните експерименти е използвана извадка от 1000 текстови ресурси (периодични издания) от дигиталната библиотека на Народна библиотека „Иван Вазов“ - Пловдив [166].
 - За валидацията на многокомпонентната формула за сходство е конструиран синтетичен набор от 60 текстови ресурси, разпределени в три групи (кълстера).
 - За поведенческа валидация е конструиран синтетичен набор от потребителски профили върху тематично сегментиран корпус. Профилите покриват шест основни случая - „студен старт“ за потребител и документ, еднотематичен интерес към една тема, „изчерпване на тема“, „смяна на контекст“ и ограничена история - и се материализират в разредена матрица „потребител-документ“ и вектор на популярност.

Верификацията на архитектурата се реализира на няколко допълващи се нива, съобразени със спецификата на дигиталните ресурси и ограниченията на наличните данни, като по този начин се осигурява пряка емпирична проверка на формулираните в предходната глава хипотези. На първо място се определя най-подходящият подход за извличане на именувани същности и се установява поточната схема за тяхната обработка. След това се извършва параметрично калибриране на метода на многокомпонентната оценка за сходство между документи, която комбинира семантична близост, тематични профили и индикатор от именувани същности; проверява се, че тази комбинирана формула превъзхожда простата семантична близост по способност да улавя действителни прилики и различия, особено в корпус от многотематични периодични издания. Качеството на матрицата на сходство и функционалността „подобни документи“ се оценяват чрез вътрешни показатели, основани на съпоставяне на извлечени характеристики: припокриване на информативна лексика, съвпадения на

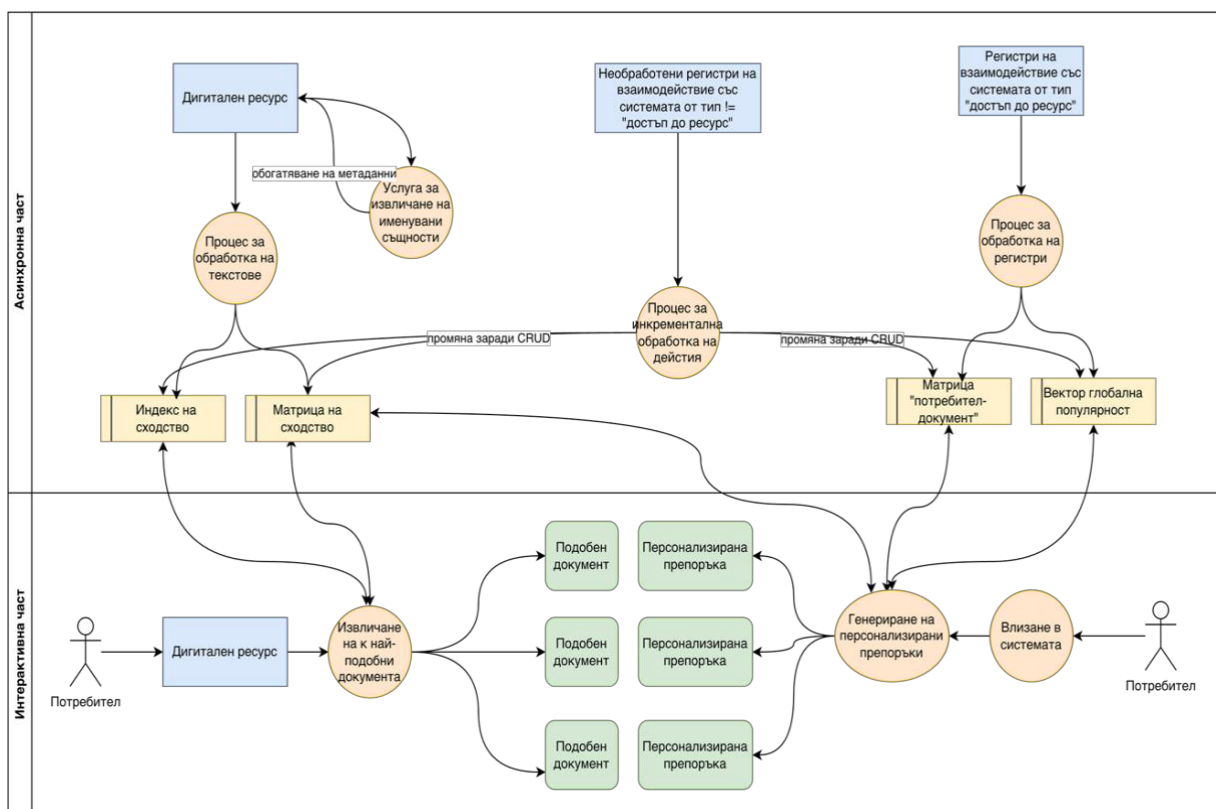
именувани същности (лица, организации, локации). Целта е да се установи, че получената структура отразява смислова близост, а не случайни съвпадения.

Паралелно се оценява модулът за персонализирани препоръки върху синтетично генерирани данни, които покриват основни и гранични сценарии: „студен старт“ (за потребители и за документи), изчерпване на интерес (затворен тематичен клъстер), динамична промяна на интересите и смесен профил. За всеки сценарий се анализира балансът между покритие и склонност към прекомерно популярни ресурси, както и устойчивостта при оскъдна история. Особен акцент се поставя върху случаите, в които хибридният алгоритъм превъзхожда класическото съвместно филтриране, базирано на елементи - например при „студен старт“ за документи и при силно разреждени поведенчески данни.

Подробните метрики, процедурите за измерване и конкретните експериментални конфигурации са изложени в следващите подраздели.

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).



Фигура 16. Архитектура на модулите за предоставяне на персонализирано съдържание

4.2.1. Асинхронен слой за подготовка и актуализация на оперативните структури

Асинхронният слой (виж фиг. 16) обединява всички изчислително интензивни процеси, чрез които корпусът от документи и регистри на взаимодействията се преобразуват в оперативни структури (предварително изчислени представяния и индекси), достъпни за бързо използване от интерактивната част. Разделянето на тежките изчисления от обслужването на заявки гарантира кратко време за отговор, предсказуемост на поведението и устойчиво мащабиране при растящи обеми и потребителска активност.

В слоя работят три независими, асинхронни компонента, съгласувани чрез обща схема от стабилни идентификатори:

1. Услуга за извличане на именувани същности.

Текстовете се обработват пакетно - извършва се унифициране (нормализиране, редукция на шум), сегментиране при дълги входове и разпознаване на лица, организации, локации и други значими обекти. Резултатът се филтрира по увереност и честота, за да се съхраняват само стабилните и информативни същности, и се публикува като

обогадени метаданни, синхронизирани с идентификаторите на документите. Така се добавя структурен семантичен показател, който може да бъде включен в оценките за близост и да подпомага обяснимостта.

2. Модул за генериране на матрица на сходство между документите.

Корпусът преминава през предварителна обработка и векторизация; използва се фрагментиране със застъпване и агрегация на представянията. Изчисляват се компоненти на близостта (глобална/усреднена, локална/по фрагменти, и тематична) и, при наличие, принос от именувани същности. Получената оценка се материализира в разредена симетрична структура, организирана като индекс на най-близките съседи (топ-k) за всеки документ, съгласуван с речника на идентификаторите. Тази структура служи пряко за „подобни документи“ и като съдържателна основа в персонализираното препоръчване.

3. Модул за взаимодействия и популярност.

Регистрите за взаимодействия със системата се филтрират до тези за достъп до ресурсите и се агрегират до разредена матрица „потребител-документ“. Паралелно се изчислява вектор на глобална популярност на документите. И двете представяния са строго подравнени към общия речник на документните идентификатори, за да позволят последователни комбинирани операции в хибридният алгоритъм.

Инкременталните обновявания се изпълняват по единна процедура, която поддържа актуалност при минимална изчислителна цена. При добавяне на документ се извличат именувани същности, извършва се предварителна обработка и векторизация, след което документът се включва в индекса по сходство (изчисляват се единствено неговите топ-k съседи), добавя се в речниците на идентификатори и се инициализира във вектора на популярност; в матрицата „потребител-документ“ се създава нова колона. При редакция се преизчисляват представянията на засегнатия документ и се актуализират съответният ред и колона в матрицата на сходство, свързаните речници и, при необходимост, свързаните метаданни; съгласуваността с матрицата „потребител-документ“ се проверява и запазва. При натрупване на нови взаимодействия се обновяват само матрицата „потребител-документ“ и векторът на популярност, без да се засягат съдържателните представяния. При изтриване документът се деактивира за интерактивната част и се премахва от индекса по сходство, речниците и вектора на популярност; колоната му в матрицата „потребител-документ“ се изтрива, за да се

предотврати появата на несъществуващи елементи в резултатите. Пълно преизчисляване се планира периодично или при съществени промени в моделите и параметрите.

4.2.2. Интерактивен слой за бързо обслужване на заявки

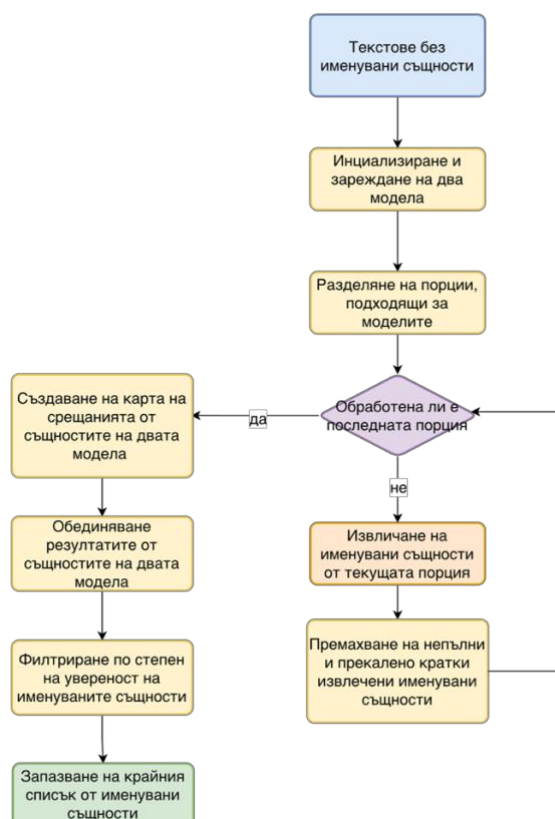
Този слой работи изцяло върху предварително изчислените оперативни структури (индекси и матрици), без да стартира тежки изчисления в момента на заявката, което осигурява ниска латентност и предсказуемо време за отговор. Поддържат се два изхода (виж фиг. 16):

- „Подобни документи“: локализира се активният документ по идентификатор и се извлича неговият ред от матрицата на сходство (или индексът на най-близките съседи); прилагат се прагове/филтри (напр. изключване на вече разглеждани материали) и се връщат топ-k резултати.
- „Персонализирани препоръки“: агрегира се сходството към документите от личната история на потребителя, добавя се корекция по вектора на популярност, вече посетените документи се изключват. При липса на история системата използва представителни (популярни) документи.

В следващите подраздели последователно са описани проектирането и реализацията на всеки компонент. Акцентът е върху основните алгоритми и използваните структури от данни, които осигуряват коректна интеграция в общата архитектура.

4.3. Услуга за извличане и структуриране на именувани същности

В рамките на архитектурата е реализирана самостоятелна услуга за извличане на именувани същности (общата последователност на обработката е показана на фиг. 17), чиято цел е да обогатява метаданни на документите със структуриран контекст (напр. лични имена, организации, географски обекти).



Фигура 17. Блок диаграма на услугата за извличане на същности

Получените списъци от същности се използват в две основни направления: (а) като допълнителен семантичен показател при конструиране на матрицата на сходство документ-документ и (б) за подпомагане на търсенето в документи. По този начин извлечените същности подпомагат едновременно прецизността и обяснимостта на препоръките, както и навигацията в колекциите.

Услугата работи в пакетен режим, извън критичния път на обслужване на потребителските заявки. Текстовете се извличат и разпределят по партии като се правят минимални трансформации, тъй като моделите за извличане на именувани същности използват пунктуацията, регистъра на буквите и дори някои често срещани стоп-думи, когато те са част от словосъчетание, като индикатор за разпознаване; следователно агресивно „почистване“ би понижило качеството на откриването.

Технологичният подход стъпва върху предварително обучени трансформър-модели, публикувани в хранилището Hugging Face [195], подбрани за добра покриваемост на български език и преносимост между среди. Проведено е предварително проучване и селекция на кандидат-модели, последвано от

експериментално тестване върху извадка от 200 документа; Таблица 9 представя сравнителните времена за обработка.

Таблица 9. Изчислителни характеристики и времена за изпълнение на моделите

<i>Модел</i>	<i>Записи / с</i>	<i>Същности на запис</i>	<i>Инициализаци я (с)</i>	<i>Изчисление (с)</i>
<i>distilbert-base-multilingual- cased-ner-hrl [196]</i>	1.03	210.72	0.05	195.08
<i>bert-base-ner-theseus-bg [197]</i>	1.24	146.59	0.02	161.18
<i>wikineural-multilingual-ner [198]</i>	0.57	160.90	0.02	352.54
<i>bert-bg-ner [199]</i>	0.71	9.50	0.01	281.43
<i>Комбинация от (1) и (2)</i>	0.57	359.27	0.01	352.30

Поради ограничението на контекстния прозорец при този клас модели (обичайно до 512 токени), дългите документи се сегментират на застъпващи се фрагменти. Макар да звучи контраинтуитивно, подаването на максимално дълъг текст като вход води до деградация едновременно на времето за обработка и на точността; затова сегментирането със застъпване гарантира, че не се губи смислов контекст на границите. Емпиричният анализ (виж таблица 10) показва най-висока пропускателна способност при 160-200 токени (≈ 1.22 документа/сек.), докато при 300+ токени се наблюдава спад (≈ 1.09 документа/сек. при 512 токени) поради квадратичната сложност на механизма за внимание. На тази основа е възприет размер *200 токени* като оптимален компромис между изчислителна ефективност и достатъчен контекст.

Таблица 10. Времена за обработка при различни размери на сегмента

<i>Размери на откъс</i>	<i>Записи / с</i>	<i>Средно закъснение (ms)</i>
128	1.2027	5.358
160	1.2188	6.596
200	1.2102	8.298
240	1.1896	10.141
300	1.1588	12.949
350	1.1176	15.691
400	1.0912	18.323
450	1.0541	21.314
480	1.0369	23.137
512	1.0896	23.42

Тестовите върху реалните данни показват, че нито един самостоятелен модел не е достатъчно устойчив. Затова е приложен двумоделен подход, при който се комбинират резултатите от два модела с най-добро съотношение между точност и бързодействие (показано на фиг. 17). Моделите се зареждат предварително (само веднъж), за да се минимизира разходът при партидната обработка. След извличането на същностите от всички фрагменти на даден документ се изгражда карта на същностите, при която: (1) се нормализира регистърът (напр. „СОФИЯ“ \equiv „София“ \equiv „софия“); (2) за всяка същност се взема максимумът по брой срещания и максимумът по показател за увереност от двата модела; (3) прилага се двупрагов филтър (минимална честота и минимална увереност - и двете конфигурируеми). Тази постобработка повишава надеждността на извличането в условия на шум и вариативност в изписването.

Резултатът се публикува във формат {документ \rightarrow {същност \rightarrow [честота, увереност]}}, като примерен запис е показан на фиг. 18. Структурата е еднозначно съгласувана с идентификаторите на документите.

```
{
  "64d36ec533d9619ac744428a": {
    "софия": [5, 0.92],
    "българия": [3, 0.88]
  }
}
```

Фигура 18. Пример за формат на per.json

В обобщение, услугата за именувани същности осигурява устойчиво, мащабируемо и обяснимо обогатяване на документите със структуриран контекст. Съчетанието от токен-сигурно сегментиране, емпирично подбран размер на сегмента (200 токени), комбиниране на два модела и строга постобработка повишава качеството на извличане в шумни и хетерогенни масиви. Периодичният пакетен режим поддържа ниска латентност на системата, а получените данни се интегрират безпроблемно както в оценката на сходство, така и в търсенето, като остават проследими и конфигурируеми.

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

4.4.1. Синонимно обогатяване на текстовото представяне (OMW-Bulgarian Wordnet)

С цел повишаване на устойчивостта на съдържателния модел спрямо лексикални вариации бе изграден синонимен речник за български език, извлечен и нормализиран от Open Multilingual WordNet (OMW) [200].

Ресурсът е приведен към леми, унифициран по регистър и графични варианти, с цел премахване на дубликати и силно многозначни форми; генериран е JSON формат „лема → списък от синоними“ (виж фиг. 19), което улеснява бързото му зареждане в асинхронния слой.

```
{
  "пояснение": [
    "обяснение",
    "разяснение"
  ]
}
```

Фигура 19. Синонимен речник

Синонимният речник се използва единствено в предварителния етап за „подобни документи“: след нормализация, лематизация и филтриране на стоп-думи към текста се добавят ограничен брой представителни синоними на съответните лемми, като това се прави преди сегментирането на дългите документи. По този начин се намалява чувствителността към повърхностни разлики в изписването и се подпомага стабилното сближаване на съдържателно близки, но лексикално различни текстове, без да се нарушава оригиналният контекст и без да се увеличава текста твърде много. Същата версия на речника се използва последователно в обработката на текста, което гарантира проследимост и възпроизводимост. Ресурсът не участва в интерактивната част и не се използва за интерфейсни обяснения; ролята му е ограничена до подобряване на векторизацията в предварителната обработка и, косвено, до повишаване на качеството на извежданите „подобни документи“.

4.4.2. Реализация на функционалния модул

Разделът представя реализацията на модула за изчисляване на сходство между текстови документи, който представлява ядрото на функционалностите за генериране на „подобни документи“ и едновременно служи като вход към модула за „персонализирани препоръки“. Целта е да се изгради надеждно и мащабируемо представяне на смисловата близост в корпуса, така че за всеки текстов документ да могат бързо и възпроизводимо да се извличат най-близките по съдържание съседи. Модулът работи в асинхронния слой на системата и е проектиран в съответствие с принципите за възпроизводимост и проследимост на изчисленията, интерпретируемост на оценките и висока изчислителна ефективност при нарастващи обеми от данни.

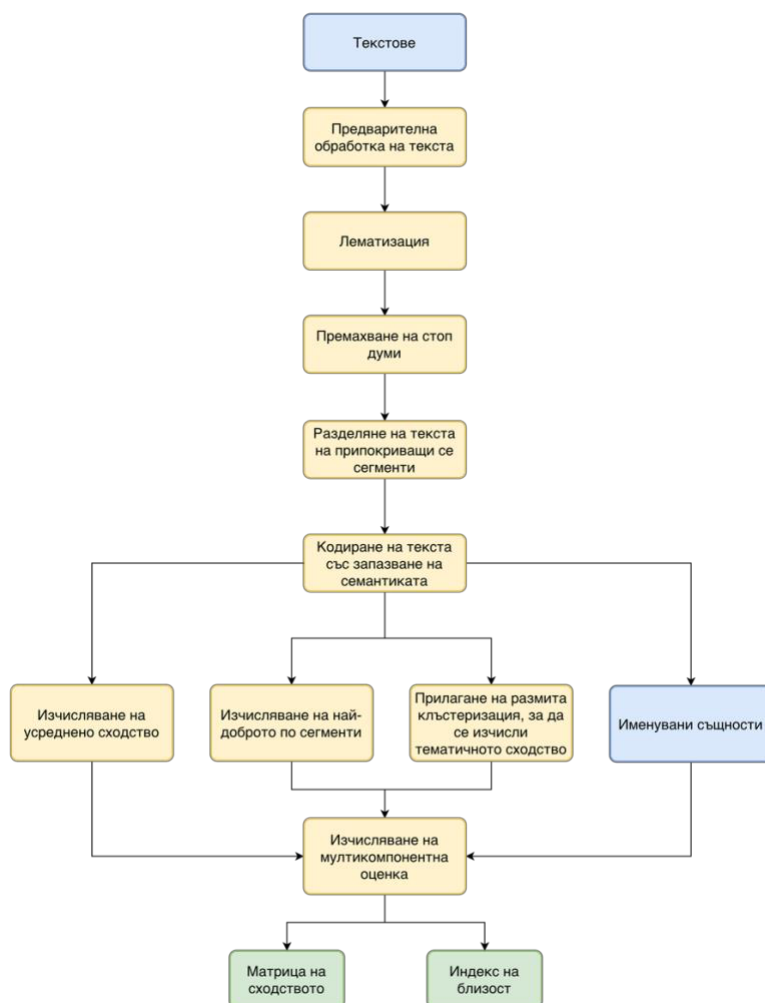
Входът включва уеднаквен текстов корпус и списъци с именувани същности за всеки документ (извлечени от отделна услуга, описана в предния раздел). Изходът е разредена, симетрична матрица на сходство документ-документ, представена като топ-k съседи за всеки ред, придружена от речник за съпоставяне `doc_id` → `row_index`. При необходимост се публикуват и спомагателни изчислени структури (например усреднени документни вектори и/или тематични профили), които подпомагат диагностика и последващи анализи.

В следващите раздели са разгледани последователността на обработката (сегментиране, векторизация, изчисляване на глобална, локална и тематична близост и, по избор, принос от именувани същности), нормализирането и параметризирането на

многокомпонентна оценка, преобразуването към разреден формат и правилата за инкрементално обновяване.

Процес по подготовка и извличане на представяния

Матрицата на сходство представлява симетрична структура с размерност $N \times N$ (където N е броят на документите, които са обработени). Процесът по генерирането на матрицата е представен на фигура 20.



Фигура 20. Процес на генериране на матрица на сходство

Той започва със зареждане на корпуса и речника на идентификаторите на документите, последвано от унифицирана подготовка на текста (нормализиране, лематизация, филтриране на стоп-думи, по избор - синонимно обогатяване). Дългите текстове се разделят на застъпващи се части по начин, съобразен с максимално допустимата дължина на входа за модела. Ако това не се спазва, моделът отрязва края на текста и се губи информация. Застъпването между частите запазва смисловата

непрекъснатост по границите, ограничава грешки от „разделяне“ на изречения и осигурява стабилни представяния за целия документ без загуби. Всеки сегмент се кодира в 384-мерен вектор с многоезичния модел MiniLM, след което на ниво документ се извеждат агрегирани представяния и, при активиран тематичен слой, профили на принадлежност, получени чрез размита клъстеризация върху сегментните вектори. Паралелно се зареждат извлечените именуванни същности и се формира разрежена матрица „документи \times същности“, която служи като структурирано входно представяне за последващи изчисления. На този етап имаме: (1) векторни представяния на документите и сегментите, (2) тематични профили, и (3) множества от именуванни същности - т.е. пълния набор от входове, върху който се стъпва при оценяването на близост.

В следващия подраздел се разглежда как е дефинирана самата оценка за сходство и правилата за обединяване на отделните компоненти (съдържателен, тематичен, по именуванни същности) в единна матрица документ-документ.

Метод на многокомпонентна оценка и конструиране на матрица на сходство

Стойността на всеки елемент на $S_{\text{финална}}(i, j)$ в матрицата на сходство отразява количествената мярка за близост между документ i и документ j в интервала $[0, 1]$. За да се осигури високо качество на оценката в условия на шумни данни (напр. OCR грешки) и изразена тематична нееднородност, характерна за периодичния печат, е разработен метод за сливане на показатели, основан на многокомпонентна оценка. Вместо да се разчита на един-единствен източник на информация, сходството между документите се моделира чрез претеглена линейна комбинация на четири независими информационни слоя, обхващащи както съдържателни характеристики на текстовете, така и информация, извлечена от именуваните същности. По този начин се повишава устойчивостта на оценката спрямо шум, тематични преходи и частично липсваща информация.

Процесът на изграждане се дефинира чрез следната математическа рамка:

1. Съдържателен показател ($S_{\text{съдържателна}}(i, j)$)

Този компонент улавя смисловата близост между текстовете, абстрахирайки се от конкретните думи (синонимия/полисемия). Тъй като вестниците и списанията са многотематични, използването само на един вектор за целия документ би довело до

загуба на информация („усредняване“ на темите). Поради това $S_{\text{съдържателна}}(i, j)$ се изчислява като линейна комбинация от три под-метрики:

- Глобално сходство ($S_{\text{усреднено}}(i, j)$): Косинусово сходство между усреднените вектори на документите. Този показател улавя общия стил и доминиращата тема на изданието.
- Локално сходство ($S_{\text{най-добро}}(i, j)$): Използва се стратегията двойна максимизация върху сегментите. Изчислява се максималното косинусово сходство между всички двойки сегменти на двата документа. Този подход е критичен за откриване на връзка между конкретна статия в един брой и аналогична статия в друг, дори ако останалата част от съдържанието е различна.
- Тематично сходство ($S_{\text{тематично}}(i, j)$): Базира се на вероятностните разпределения, генерирани от алгоритъма за размито клъстеризиране. Сходството се измерва чрез скалярно произведение на векторите на принадлежност към темите. Това позволява свързване на документи, които споделят комбинация от теми (напр. 30% Политика и 20% Икономика).

2. Показател от именуваните същности ($S_{\text{именувани същности}}(i, j)$)

Семантичният компонент понякога пропуска връзки, базирани на редки лични имена или специфични географски понятия. За да се компенсира това, се въвежда втори слой на сходство, базиран на именуваните същности. За всеки текстов документ е извлечено множество от същности. Сходството между два документа се оценява според припокриването на тези множества. По този начин съвпаденията по именувани същности предоставят ясен и надежден фактологичен показател, който допълва общата картина за тематична близост.

3. Финална оценка ($S_{\text{финална}}(i, j)$)

Крайната оценка на сходство между текстовите документи се получава чрез линейна комбинация между семантичния (1) и фактологичния слой (2). А параметърът, оказващ с каква тежест да се ползват именуваните същности, играе ролята на регулатор, който позволява експериментално настройване на баланса между семантичното и фактологичното сходство. Целта на този подход е да се изследва дали даването на лек превес на съвпаденията по именувани същности (имена, места) би могло да действа

като стабилизиращ фактор за семантичните модели, повишавайки релевантността на препоръките за крайния потребител.

Обосновката за включването на всеки от четирите информационни слоя, подкрепена с практически сценарии от корпуса, е представена детайлно в Таблица 11.

Таблица 11. Роля на компонентите в многокомпонентна крайна оценка

<i>Компонент</i>	<i>Проблем, който решава</i>	<i>Пример</i>
$S_{\text{усреднено}}(i, j)$ <i>косинусова близост</i>	Общ стил и език. Улавя глобалния тон на изданието.	Свързва два броя на едно периодично издание просто защото стилът на писане е еднакъв.
$S_{\text{най-добро}}(i, j)$ <i>косинусова близост по сегменти</i>	Локално съвпадение. Открива конкретна статия вътре във вестника.	Ако в Брой 5 и Брой 100 има статии за "АЕЦ Белене", този компонент ще ги свърже, дори останалата част на вестниците да е различна.
$S_{\text{тематично}}(i, j)$ <i>алгоритъм на размити средни</i>	Тематичен микс. Моделира разнородността на съдържанието.	Позволява документът да бъде намерен и при търсене за "Политика", и при търсене за "Култура".
<i>Именувани същности</i>	Фактология. Свързва документите чрез конкретни личности/места.	Свързва фейлетон за "Ахмед Доган" с новина за "Ахмед Доган", дори ако в единия текст се говори иронично, а в другия сериозно (семантиката е различна, но лицето е същото).

Получената оценка и по-точно матрица се материализира в компактни структури за бърз достъп. По-долу са описани форматите на съхранение и механизмите за тяхната актуализация.

Исходни оперативни структури: формати, ефективност на съхранение и актуализация

Ефективността на препоръчващата система в реално време зависи не само от точността на математическите модели, но и от начина на съхранение и достъп до генерираното знание. За да се осигури минимална латентност при обслужване на потребителските заявки, архитектурата разделя процеса на два слоя: изчислителен етап и интерактивна част.

Както вече беше споменато, генерирането на матрицата на сходство се прави в асинхронния слой (изчислителния етап). Получената матрица се съхранява в двоичен формат (.npy), което позволява мигновеното ѝ зареждане в оперативната памет и осигурява висока производителност на модула за препоръчване в реално време. Съхраняват се също речници за съпоставяне на идентификатори, усреднените векторни представяния и тематични профили (когато са активирани). Използването на двоичен формат позволява прилагането на техниката за „изобразяване в паметта“, осигурявайки директен достъп до данните без необходимост от пълното им зареждане в оперативната памет, което е ключово за мащабируемостта на системата.

Тъй като матричните операции работят с целочислени индекси, а информационната система оперира с уникални идентификатори на обектите, архитектурата поддържа и специализиран индекс за трансляции. Тази оперативна структура реализира двупосочно съответствие между системните идентификатори и позицията на документите в матрицата, позволявайки адресна трансляция с константна сложност $O(1)$.

Данните се сериализират в компактни формати, оптимизирани за бърз достъп и малък ресурсен отпечатък:

- `sim_hybrid.npy` - финалната матрица на сходство ($N \times N$, ~1 MB за ~1000 документа, `float32`), която агрегира съдържателните и фактологичните компоненти и позволява директно извличане на ред с константна сложност;
- `item_index.json` - речник за двупосочна трансляция „doc_id ↔ матричен индекс“ (~17 KB за ~1000 документа), използван за незабавно локализиране на документ в матрицата;

- `mean_embs.npy` - усреднени семантични вектори (384 измерения) в `float16` (~375 KB за ~1000 документа), пригодни за диагностика и за динамични оценки при нови документи;
- при активиран тематичен слой се добавят `topic_profiles.npy` (~39 KB за ~1000 документа) с векторите на принадлежности към теми и `fuzzy_centers.npy` (~60 KB за ~1000 документа) с центроидите на тематичните клъстери.
- за нужди на дълбок анализ и отстраняване на проблеми се съхранява и `chunk_embs.pkl` (~36 MB за ~1000 документа) със суровите вектори на всички сегменти; този файл не е необходим за ежедневната експлоатация.

Целият конвейер е оптимизиран за хардуерно ускорение (CUDA/MPS), партидна обработка и инкрементално допълване на колекциите: изчисленията работят на партиди, векторите се записват в полу-прецизност (`float16`) там, където това не влияе на точността на косинусовото сходство, а сериализираните структури позволяват последователно добавяне/актуализиране без пълно преизчисляване. Тази организация гарантира бързо зареждане, ниска латентност при заявки и възпроизводимост на резултатите.

Актуалността се поддържа чрез инкрементални процедури: при добавяне на документ се изчисляват представяния само за новия елемент и се актуализират релевантните редове/колони в индекса на сходство; при редакция се преизчисляват засегнатите представяния и зависимости; при изтриване ресурсът се деактивира за извличане и се премахва от индексите.

Поради предварителното изчисление интерактивното извличането на „подобни документи“ се свежда до четене на съответния ред от матрицата, филтриране и подреждане по стойности.

Извличане на „подобни документи“

Самият процес на препоръчване се инициира в момента на достъпване на конкретен документ в дигиталната библиотека. Алгоритъмът за извличане на „подобни документи“ следва логиката на метода на k -най-близките съседи (k -NN). Първоначално, идентификаторът на текущия документ се транслира до съответния матричен индекс. Впоследствие системата извлича директно съответния ред от матрицата на сходство, който съдържа количествените оценки за близост спрямо всички останали документи.

Полученият вектор се подлага на сортиране в низходящ ред, като се прилагат правила за филтрация и отрязване на резултати под дефиниран праг на увереност.

В резултат се селектират първите k записа с най-висок коефициент на сходство. Техните матрични индекси се преобразуват обратно в системни идентификатори чрез транслационния индекс, след което се извличат необходимите данни за визуализация. Този подход гарантира, че сложните семантични и тематични изчисления са изнесени в етапа на предварителната обработка, а потребителят получава мигновен отговор, базиран на ефективни операции за четене и сортиране.

4.4.3. Експериментална валидация на функционален модул „подобни документи“

За да се оцени качеството на предложената архитектура - и в частност на матрицата на сходство между документите - както и да се определят оптималните стойности на тегловните коефициенти, участващи в многокомпонентната оценка (α , β , γ , λ), следва да се преодолеят няколко съществени затруднения: липса на предварително известни сходства в реалния корпус, голям брой документи и значителна дължина на отделните текстове. Поради това е формулирана многоетапна експериментална процедура, при която валидирането се извършва чрез съчетаване на автоматизирана оптимизация и качествен анализ върху представителни извадки с нарастващ обем.

Първият етап представлява претърсване на пространство от параметри по предварително зададена решетка от стойности за параметрите (α , β , γ , λ), съпроводено от компонентно изключване (Ablation study - постъпателно изключване на отделни съставки на оценката - глобална семантика, локални съвпадения, тематични профили, именувани същности), за да се измери пределният принос на всеки слой. Оценяването се извършва върху реални извадки с контролирани „истинни“ връзки, както и чрез вътрешни метрики: корелация спрямо базови модели, припокриване на най-близките k съседи и устойчивост на резултатите при повторно вземане на подизвадки.

Вторият етап е цялостна оценка на метода на многокомпонентна оценка, където се анализират: (1) вътрешна кохезия в тематично еднородни групи, (2) разграничимост между тематично несвързани подкорпуси, (3) улавяне на частични (междутематични) връзки и (4) стабилност на подреждането при инкрементални обновявания. Резултатите се интерпретират едновременно количествено и качествено, което позволява аргументиран избор на параметри и прозрачно обвързване на архитектурните решения с наблюдаваната ефективност.

Валидация и калибриране на тегловите коефициенти

Експерименталната постановка се базира на хипотезата, че семантично свързаните документи следва да имат висока степен на припокриване не само на ниво „векторно пространство“, но и на ниво именувани същности, термини и дори отделни думи. Това също означава, че многокомпонентна оценка за сходство, която включва глобална семантика, локални съвпадения, тематични профили, именувани същности ще генерира по-добри препоръки за „подобни документи“. За да се провери тази хипотеза и да се оптимизират параметрите, бе проведен серия от контролирани експерименти, целящи да установят баланса между четирите компонента на сходство ($S_{\text{усреднено}}(i, j)$, $S_{\text{най-добро}}(i, j)$, $S_{\text{тематично}}(i, j)$) и $S_{\text{именувани същности}}(i, j)$), който максимизира обективната информационна свързаност между препоръчаните документи.

В този раздел се представят резултатите от параметрична оптимизация (решетка за претърсване, Grid Search) върху контролирани тестови извадки с различен обем. За целта обаче преди оптимизацията е конструирана матрица („златен стандарт“), която комбинира: (а) процент на припокриване на лексикални единици (след лематизация и синонимно разширяване) и (б) припокриване на именувани същности. Тя служи за референтна стойност, спрямо която се измерва степента на приближение на изчислената матрица на сходство при дадени параметри.

За да се провери стабилността на метода и чувствителността върху извадки, е следван итеративен протокол с постепенно нарастване на обема и разнообразието на данните:

- Малки извадки: 5 групи \times 200 документа - чувствителност към локални тематични вариации.
- Средни извадки: 2 групи \times 500 документа - баланс спрямо статистическия шум.
- Контролна извадка: 1 група \times 1000 документа - проверка на мащабируемостта.

За всяка група се изпълняват последователно:

1. Оптимизация на параметрите чрез решетка за претърсване с цел максимално приближение до златния стандарт;
2. Генериране на модела - изчисляване на матрица на сходство при намерените стойности;
3. Оценка на качеството чрез независими проверки, описани по-долу.

Оценяването се прави на три допълващи се класа проверки:

- а. Структурна верификация. Проверяват се базови свойства на матрицата: диагоналните елементи да са близо до 1.0 (идентичност), а извъндиагоналното разпределение да показва разумна вариативност.
- б. Лексикално сходство. Измерва се припокриването на лексикален запас между сродни документи след лематизация и премахване на често срещана служебна лексика; включва се и синонимен речник, така че да се отчита семантична близост при различна семантична форма (напр. „автомобил“/„кола“).
- в. Семантична свързаност. Валидира се споделеност на ключови именувани същности (лица, организации, географски обекти) между референтния и препоръчаните документи, което е силен индикатор за тематична близост в контекста на библиотечни корпуси.

Тази комбинация от „златния стандарт“, контролирани извадки и многостепенна оценка позволява едновременно параметрична оптимизация и надеждна проверка на качеството и устойчивостта на модела при различни мащаби на данните.

Експерименталните резултати от оптимизацията показаха, че оптималната стойност за теглото на именуваните същности е $\lambda = 0.5$. Този резултат дефинира архитектурата като симетричен хибрид, при който семантичният анализ и фактологичното съвпадение имат равнозначна роля при формирането на препоръката. Тази относително висока стойност (в сравнение със стандартните 0.1 - 0.2 при общи текстове) се обосновава със спецификата на дигиталната библиотека като информационна система, която съдържа много периодични издания. За разлика от художествената литература, където стилът и емоцията са водещи, в периодичния печат и архивните документи информационната стойност е концентрирана около конкретните именувани същности (лица, организации и локации).

Стойността $\lambda = 0.5$ действа като ефективен регулатор, който:

- Потиска „мними“ семантични връзки между звучно близки, но фактологично различни текстове и приоритизира документи, които споделят общ контекст по лица, организации и локации.

- Гарантира, че в списъка с препоръки ще попаднат документи, които споделят общ фактологичен контекст, което е основният потребителски интерес при работа с периодични издания.

Експерименталните тестове за оптимални стойности на останалите параметрите α , β , γ , на които се базира многокомпонентна оценка (изведена в раздел 3.2.4), показаха $\alpha = 0.47$, $\beta = 0.07$ и $\gamma = 0.47$, което показват, че общият смисъл и тематичната категория са водещи, докато съвпадението на отделни откъси има второстепенно значение.

Това разпределение води до два основни извода:

1. Доказана полза от тематичното моделиране. Високата стойност на параметъра $\gamma = 0.47$ оправдава добавянето на изчислителна сложност при добавянето на подход за клъстеризация и служи за емпирично потвърждение за ефективността на компонента за размита клъстеризация. Тъй като вестниците съдържат разнородни теми (напр. политика, спорт и култура в един брой), тяхното тематично разпределение се оказва също толкова важно за групирането им, колкото и усреднения семантичен вектор α . Това показва, че решението да се използва е било архитектурно правилно.
2. Потискане на шума от частични съвпадения - ниската тежест на параметъра β (0.07) индикира, че наличието на единични сходни пасажки не е достатъчно, за да се определят два документа като „подобни“. В контекста на периодичния печат това е полезно поведение, тъй като предотвратява грешни препоръки, базирани на повтарящи се реклами, обяви или стандартни рубрики, които нямат съществена информационна стойност.

Паралелно е проведен и аблационен анализ (виж табл. 12), който систематично оценява приноса на всеки компонент. Наблюдава се устойчива динамика на метриките в три последователни фази:

- Фаза на генерализация (базова конфигурация с $S_{\text{усреднено}}(i, j)$) - усредненият семантичен вектор показва високи стойности по лексикални метрики поради склонност към обобщаване и намиране на множество „широки“ асоциации - полезно за обхват, но с риск от семантичен шум.

- Фаза на спецификация (добавяне на $S_{\text{най-добро}}(i, j)$ и $S_{\text{тематично}}(i, j)$) - въвеждането на локални съвпадения по фрагменти и тематични профили води до отчетлив спад на чисто лексикалните показатели, но това е очакван ефект от повишена селективност - филтрират се повърхностни прилики и се запазват по-консистентни концептуални връзки.
- Фаза на контекстуализация (с добавяне на $S_{\text{именувани същности}}(i, j)$, при $\lambda = 0,5$) - фактологичният слой възстановява част от „изгубените“ кандидати, но само когато има споделен контекст по същности, така че финалният баланс съчетава прецизност (намален шум) и обяснимост (ясни основания „по общи същности“).

Таблица 12. Аблационен анализ

<i>Размер на извадката</i>	<i>Усреднено</i>	<i>Усреднено, най-добро и тематично</i>	<i>Многокомпонентна (и именувани същности)</i>
200	0,1022	0,0712	0,0959
200	0,1109	0,0927	0,1061
200	0,0884	0,0831	0,0918
200	0,1021	0,08	0,942
200	0,0937	0,0665	0,0927
500	0,1067	0,0808	0,0954
500	0,0984	0,0772	0,0911
1000	0,1025	0,0776	0,0895

В обобщение, проведеното декомпозиране на модела (Ablation Study) през всички тестови сценарии (малки, средни и големи корпуси) разкрива устойчива и характерна тенденция в поведението на метриките за качество. Наблюдава се специфичен профил на резултатите, при който добавянето на сложност първоначално води до спад в метриката за лексикално припокриване, последван от значително възстановяване при

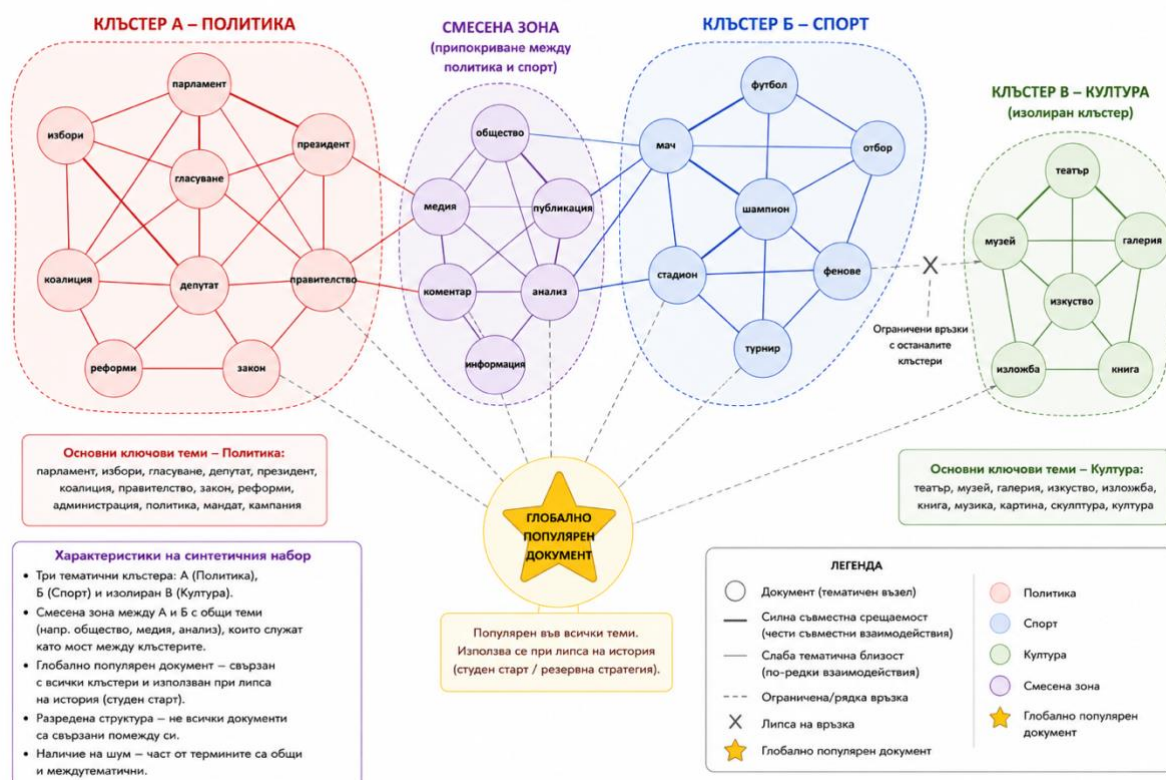
финалната интеграция. Това означава, че базовият модел максимизира обхвата, докато метода на многокомпонентната оценка повишава прецизността и устойчивостта; избраната конфигурация с равностойни приноси на семантичен и фактологичен слой е най-подходяща за периодични многотематични издания в дигитални библиотеки, където надеждността и интерпретируемостта на връзките са водещи.

В контекста на настоящата работа аблационният анализ се използва като подготвителна стъпка към цялостната експериментална валидация: първо се оценява самостоятелният принос на всеки компонент (усреднено семантично сходство, локални съвпадения по фрагменти, тематични профили, именувани същности), след което резултатите от този анализ насочват калибрирането на тегловните коефициенти в комбиниранията оценка. На тази база следващият раздел представя систематичните тестове на конфигурацията и ефекта от взаимодействието между отделните слоеве върху качеството на препоръките.

Експериментална валидация на метода на многокомпонентна оценка за „подобни документи“

В предходния раздел бяха калибрирани тегловните коефициенти α , β , γ , λ на метода на многокомпонентната оценка и бе установена работна конфигурация на модела. В настоящия етап тези стойности се използват като фиксирани параметри, за да се оцени емпирично приносът на всеки компонент към крайния резултат и да се провери до каква степен включването на тематичен слой и именувани същности повишава качеството на сходствата.

С цел контролируемо и възпроизводимо тестване е конструиран синтетичен корпус, който позволява прецизно управление на тематиката, нивото на общ речников шум и наличието на именувани същности. Корпусът съдържа 60 документа, организирани в три равностойни групи: (А) „Вестници“ (политика + спорт) с асоциирани същности „Иван Иванов“ и „Христо Стоичков“; (Б) „Списания“ (готварство) с „Ути Бъчваров“; (В) „Бюлетини“ (политика) с „Иван Иванов“. За да може синтетичните данни да се доближават до реални данни във всеки документ е добавен значим дял обща лексика, така че задачата да не се свежда до тривиално лексикално припокриване. „Златният стандарт“ за сходство е дефиниран детерминистично: пълно сходство (1.0) за двойки в рамките на една и съща група; частично (0.7) между „Вестници“ (P+S) и „Бюлетини“ (P) поради споделения политически компонент; и нулево (0.0) за несвързани двойки. Различните теми в клъстерите и отделните клъстери са онагледени на фигура 21.



Фигура 21. Тематични клъстери и връзки между документите в синтетичния набор от данни

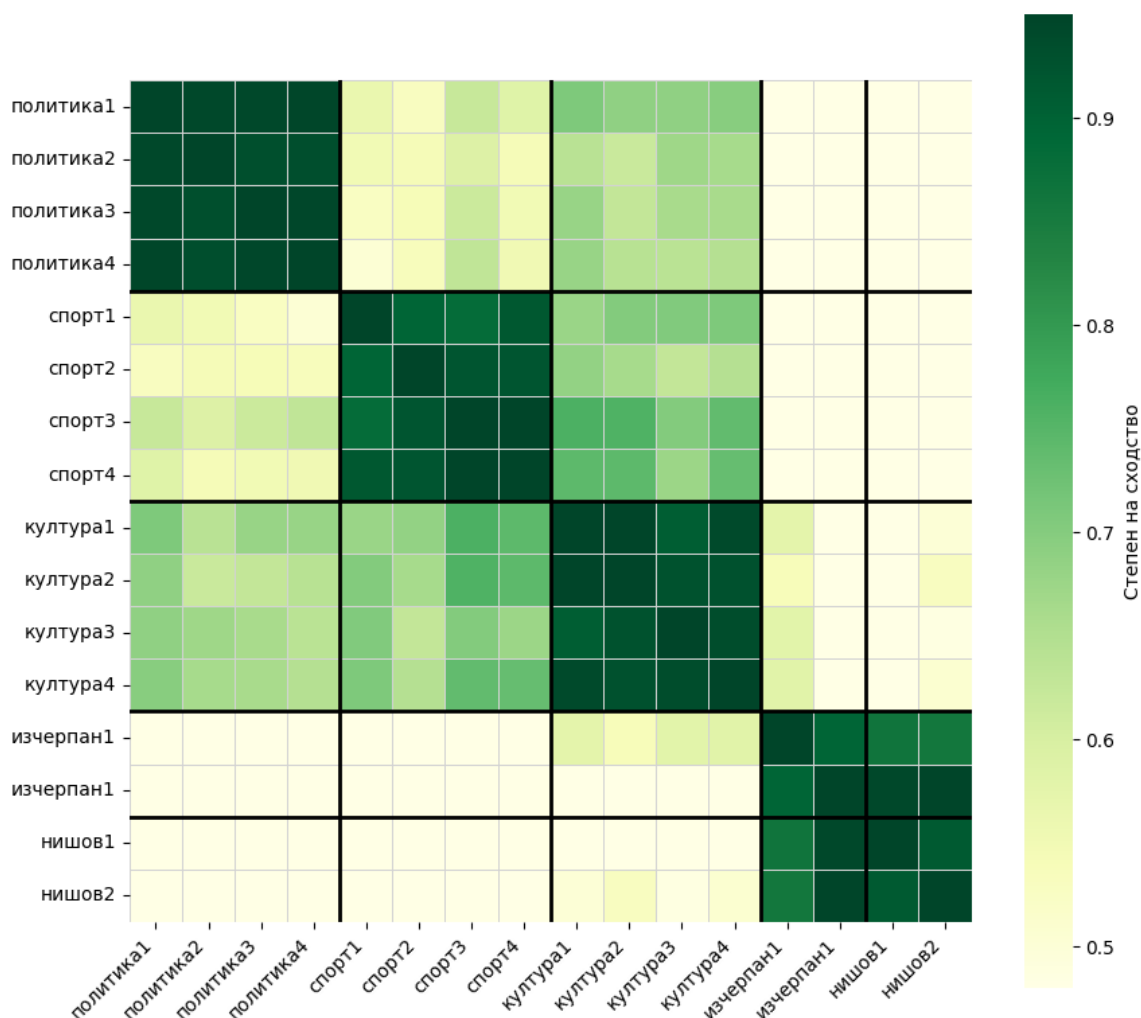
Върху този корпус се изпълняват три постановки с вече фиксираните тегла:

- Базова конфигурация само с усреднен семантичен вектор ($S_{\text{усреднено}}(i, j)$)
- Съдържателна конфигурация, която комбинира глобално, локално и тематично сходство ($S_{\text{усреднено}}(i, j)$, $S_{\text{най-добро}}(i, j)$ и $S_{\text{тематично}}(i, j)$)
- Всички компоненти на многокомпонентната оценка (добавя се и $S_{\text{именовани същности}}(i, j)$)

Оценката се извършва по три допълващи се направления: вътрешна кохезия (способност за обединяване на документи от една и съща тематична група), дискриминация между несвързани теми (ограничаване на фалшиви сходства, породени от обща терминология) и откриване на частични връзки между документи с частично тематично припокриване.

Получените резултати очертават последователна картина. Базовата конфигурация (само косинусова близост между документите), представена на фигура 22 с част от документите, показва висока чувствителност към широки семантични зависимости

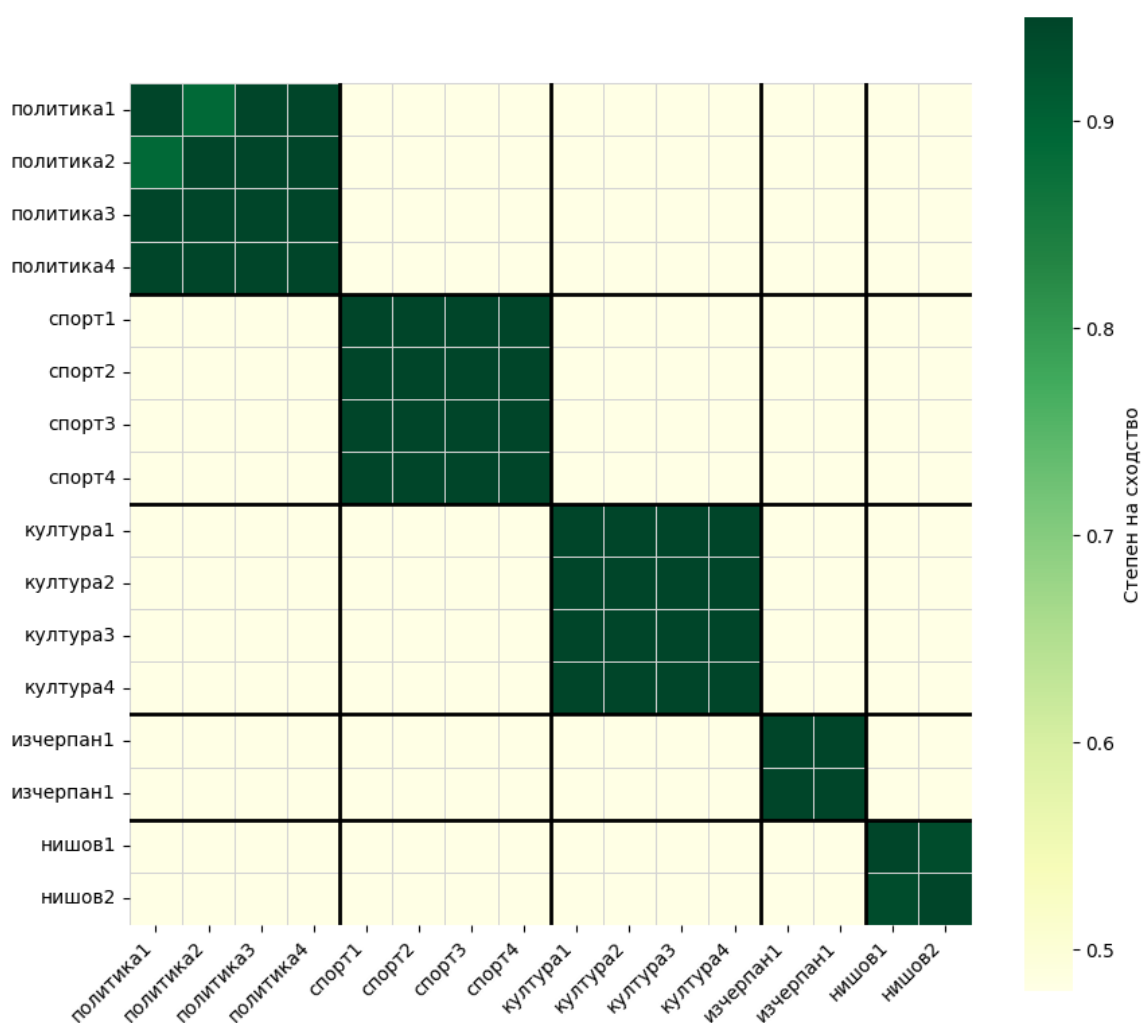
между документите. Съдържателната оценка, базирана единствено на косинусова близост между семантичните представяния, формира ясно разграничени тематични клъстери, но същевременно допуска и по-широки междутематични връзки. Това поведение е очаквано, тъй като сходството се определя предимно от локални текстови зависимости и общ контекст между документите.



Фигура 22. Топлинна карта на сходството между документите при използване единствено на косинусова близост

Резултатът от тестовете, когато се ползва предложеният метод на многокомпонентна оценка, представен на фигура 23, показва по-плавна и по-свързана структура на сходството между документите. Освен ясно обособените тематични клъстери, се наблюдават и преходни области между отделните групи, което показва наличие на по-широки семантични зависимости. Многокомпонентната оценка

допълнително укрепва семантично релевантните връзки чрез фактори като тематична близост и намалява влиянието на случайни съвпадения, причинени от обща терминология. В резултат на това когато се ползва многокомпонентната оценка се наблюдава по-ясно разграничаване на тематичните клъстери и ограничаване на междутематичните шумови зависимости спрямо базовата конфигурация.



Фигура 23. Топлинна карта на сходството между документите при многокомпонентна оценка

Наблюдава се също, че при многокомпонентната оценка се запазват ограничени връзки и между по-слабо свързаните групи, като дори документите със смесени характеристики формират умерени връзки с повече от една тематична област. Това поведение е особено важно при работа с разредени библиотечни данни и многотематични периодични издания. В тези случаи комбинирането на няколко

независими компонента позволява изграждане на по-устойчива структура на сходство и по-стабилни препоръки.

За по-конкретна илюстрация на поведението на двата подхода Таблица 13 представя примерни препоръки, генерирани за част от документите в синтетичния набор от данни. Целта е да се демонстрира как различните механизми за оценка на сходството влияят върху формирането на връзки между документите, особено при специализирани и преходни случаи.

Таблица 13. Примерни препоръки, съответстващи на експерименталния набор

<i>Документ</i>	<i>Базова конфигурация (косинусова близост)</i>	<i>Многокомпонентна оценка</i>
<i>политика</i>	политика и култура	политика
<i>спорт</i>	спорт и култура	култура
<i>култура</i>	политика, спорт, култура	култура
<i>нишов</i>	нишов и изчерпан	нишов
<i>изчерпан</i>	нишов и изчерпан	изчерпан

Особено показателни са случаите със специализираните и преходните документи, при които многокомпонентната оценка формира по-устойчиви връзки в сравнение с базовата конфигурация. Докато при косинусовата близост препоръките остават концентрирани основно в рамките на локални тематични групи, предложеният подход успява да установява допълнителни зависимости между тематично близки ресурси със сходен контекст, но ограничено директно текстово припокриване.

Компромисът при предложеният подход е по-сдържаното разпознаване на слабо изразени частични сходства. В контекста на академичните и библиотечните системи обаче това поведение е приемливо и дори желано, тъй като приоритет се дава на надеждните и обясними връзки между документите, а не на случайни асоциации, породени от обща терминология или шум в данните.. *Тези резултати потвърждават*


хипотезата формирана в 3.2.4, че методът на многокомпонентна оценка за сходство спомага за генерирането на по-устойчиви и полезни препоръки.

Практическа демонстрация на модула „подобни документи“

За да изобразим примерен сценарий на употреба, бе създадено демо приложение, реализирано с помощта на Streamlit, което показва идентифицираните като „подобни“ документи спрямо зададен източник (виж фиг. 24).


За целите на обяснимостта числовите коефициенти на сходство не се показват директно, а се отнасят към предварително зададени категории по следната работна скала: 0.85 - 1.00 - „идентичност“ (почти дубликати/препечатки); 0.65 - 0.85 - „високо сходство“ (обща тема и споделени ключови обекти); 0.40 - 0.65 - „свързани“ (тематична близост при различни детайли); 0.00 - 0.40 - „слаба връзка“ (типично се филтрират от интерфейса). Тази стъпка повишава прозрачността за потребителя („защо ми се предлага този документ“) и улеснява валидирането на алгоритъма без да се изисква интерпретация на абсолютни стойности. Праговете са емпирично определени върху текущия корпус и подлежат на калибриране при промяна на данните, конфигурацията на модела или избраната методика за оценка.

Подобни документи

 **Източник**

ID: 62eaa3c2f8e70b632213100d

/ ПЛОВДИВСКИЯ КИРИЛЪ Да вършимъ добро съки човѣкъ иска да получава само добрини отъ Бога и отъ хората, като че ли само той е тукъ на земята и само той има правото да получава. Дойде ли редъ, обаче, той да стори нѣкакво добро, тегли се, колебае се, пъкъ най-сетне ще се реши да го направи, или да не го направи. А иначе трѣбва да блде: както искаме да ни правятъ добро и съ удовольствие приемаме добринитъ, които ид- ватъ отъ Бога и отъ хората, тъй трѣбва всѣкога, и безъ коле бание да сме готови и ние да вършимъ добрини, понеже и дру- гитъ човѣци, подобно на насъ, се считатъ въ п...

 **Подобни**

62eaa3c2f8e70b6322130fe6

Високо сходство

Абонаментъ за година 15 лева. Нѣко отъ едно мѣсто се записватъ за списанието най-малко 10 души и се праща въ обща връзка...

62eaa3c2f8e70b6322130ff0

Свързан

Год. IX. Пловдивъ, мартъ 1941 год. Брой 3. Абонаментъ за година 15 лева. Яко отъ едно мѣсто се записватъ за списанието ...

62eaa3c2f8e70b6322130fbd

Свързан

ГОД. V. пловдивъ, МАРТЪ 1937. ГОД. БРОЙ 3. Жетвата е голѣма, а работницата " "малко (Лука 10:2). я.7шъавяган кдияявтс...

62eaa3c2f8e70b6322130fd9

Свързан

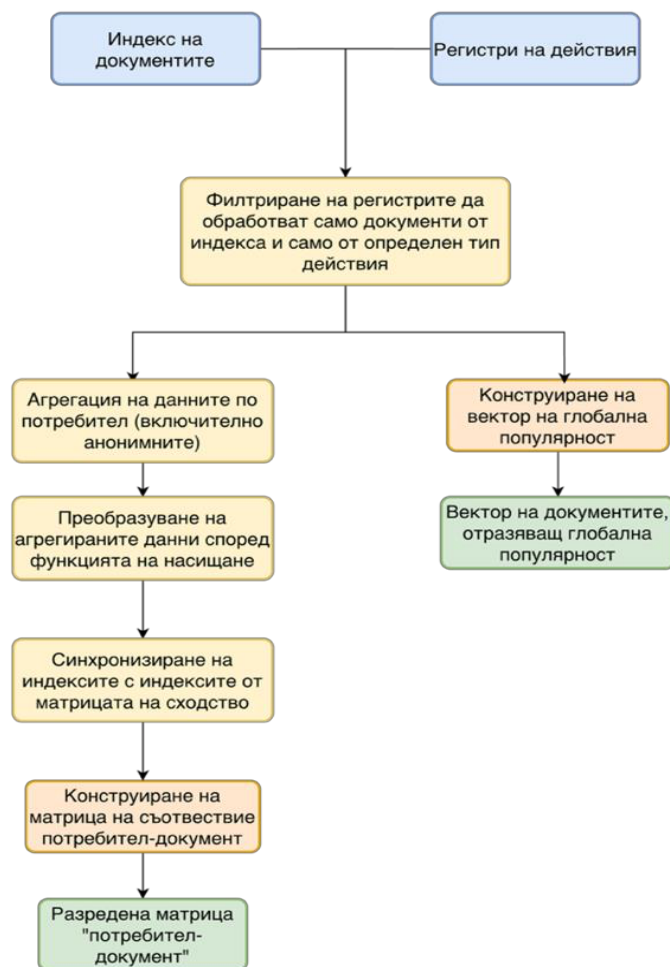
Излиза единъ. 10ктъ въ месеца, освенъ юлий и августъ АБОНАМЕНЪ ЗА ГОДИНА.....15 ЛЕВА (Я...

Фигура 24. Практическа демонстрация на „подобни документи“

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

4.5.1. Реализация на модула

Модулът за генериране на разредената матрица „потребител-документ“ има за цел да превърне регистрите на взаимодействията със системата във формално, съгласувано и ефективно за изчисления представяне на предпочитанията, върху което да стъпва персонализираното препоръчване. Като вход се използват само регистрите за достъп до информационните ресурси (псевдонимизиран идентификатор на потребител, идентификатор на документ, времеви печат и тип действие), съгласувани с общата схема за идентификатори на документите. Като изход модулът генерира (1) разредена матрица „потребител-документ“, чиито стойности са имплицитни тегла, получени от насищаща се трансформация на валидираните преглеждания, и (2) вектор на глобална популярност на документите, нормализиран в $[0, 1]$. Поддържат се и необходимите речници за еднозначно съпоставяне между системните идентификатори и матричните индекси, така че поведенческите данни да се комбинират последователно със съдържателните показатели. Тази конфигурация осигурява едновременно ниска латентност при обслужване (четене и сортиране върху разредени структури) и предвидима поддръжка при редовни инкрементални обновявания. Пълният процес е онагледен на фиг. 25.



Фигура 25. Генериране на разредена матрица „потребител-документ“

Конструиране на разредена матрица „потребител-документ“ и претеглена стойност на увереност на препоръките

В процеса на изграждане на архитектурата се разграничават два основни модула. Първият модул включва генерирането на матрицата на сходство, която моделира връзките между документите на база тяхното съдържание. По линията на персонализираното препоръчване вторият модул на системата обработва регистрите на взаимодействията и превръща наблюдаваното поведение в формализирани структури за изчисление.

Поради липсата на експлицитни оценки и детайлни поведенчески показатели (напр. време на престой или дълбочина на преглед), всеки регистриран достъп се тълкува като имплицитна положителна оценка за съответния документ. Обаче броят на отваряния на един документ не е линейна мярка за интерес. Например, десет отваряния на един документ не означават непременно десет пъти по-голям интерес от едно отваряне. Често,

тези повторения могат да бъдат резултат от навигационен шум или презареждане на страницата. За да се справи с този проблем, в системата е въведена функция на насищане (Saturation Function), която преобразува суровия брой отваряния на един и същи информационен ресурс ($c_{u,i}$) в претеглена стойност на увереност (формулата е дефинирана в 3.3.2). Формулата гарантира, че първоначалният преглед на документ ще има базова стойност от 1.0, като всяко следващо посещение увеличава увереността в интереса към документа с намаляваща стъпка. Стойността на увереността обаче се ограничава до максимум 2.0, което предотвратява изкривяването на модела от аномални поведения, като например многократни автоматизирани достъпи до ресурса.

След филтриране на нерелевантни събития и агрегация по потребител и документ се формира разрежена матрица „потребител-документ“ $W \in \mathbb{R}^{U \times I}$ със стойности $w_{u,i}$ единствено там, където е наблюдавано взаимодействие. Поради голямата разреженост (всеки потребител консумира нищожна част от корпуса), W се съхранява във формат Compressed Sparse Row (CSR) [201], което пази само ненулевите елементи и координатите им и позволява ефективни матрични операции и мащабируемост върху стандартен хардуер.

Успоредно с W се изчислява вектор на глобална популярност $p \in \mathbb{R}^I$, който отразява агрегирания интерес към всеки документ на ниво система. Той се дефинира като сума от имплицитните тежести по потребители, $\sum_u w_{u,i}$, и по избор може да се нормализира в $[0, 1]$, за да е съпоставим с останалите компоненти на модела. Този вектор служи като стабилизиращ индикатор при оскъдна или липсваща индивидуална история (напр. нови или анонимни потребители) и се използва в хибридната алгоритъм при персонализираното подреждане.

Изходът от модула включва три основни оперативни структури: разредената матрица „потребител-документ“, вектор на глобалната популярност на документите и речник за съпоставяне на потребителски идентификатори към редове (user_index.json), съгласувани с общия индекс на документите (item_index.json). Те се публикуват във компактни формати, позволяващи бързо зареждане и инкрементални обновявания. В следващия раздел се представят техните формати, начини на съхранение и процедури за актуализация.

Исходни оперативни структури: формати, ефективност на съхранение и актуализация

Резултатът от обработката на данните е матрица на взаимодействията, която има размерности $U \times I$, където U е броят на потребителите, а I е броят на документите в дигиталната библиотеката или поне тези, които са обработени до момента. За да се гарантира съвместимост със съдържателния модул, се налага пространствена синхронизация спрямо общия индекс на документите (`item_index.json`), изграден при конструирането на матрицата на сходство: j -тата колона в матрицата на взаимодействията еднозначно съответства на j -тия ред/колона в матрицата на сходство, което осигурява съвместимост между двете матрици за бъдещи алгебрични операции, използвани при хибридното генериране на препоръки.

Тъй като само малка част от документи се разглеждат от всеки потребител, матрицата на взаимодействията е силно разрежена, като повече от 99% от елементите са нулеви. За да се оптимизира паметта и да се подобри ефективността на системата, матрицата се съхранява в разрежен CSR формат (Compressed Sparse Row) [201], използвайки библиотеката SciPy [193]. Този формат съхранява само ненулевите стойности и техните координати, което позволява на системата да обработва милиони потребители и документи върху стандартен хардуер.

Исходът на модула се публикува като компактни файлови структури, достъпни за бързо зареждане от интерактивната част:

- `user_item_interactions.npz` - разредената матрица „потребител-документ“ в компресиран двоичен формат;
- `user_index.json` - съпоставяне между потребителски идентификатори и редове в матрицата;
- `doc_popularity.json` - агрегирана популярност по документ, използвана като стабилизиращ индикатор при оскъдна или липсваща история.

Тези структури гарантират кратко време за отговор и последователност при комбиниране на поведенчески и съдържателни показатели.

Инкременталните обновявания обхващат добавяне на нови записи от регистрите, корекции и изтривания. При постъпване на нови взаимодействия се актуализират съответните редове в матрицата „потребител-документ“ и компонентите на вектора на популярност, без да се изисква пълно преизчисление. При изтриване или промяна на

документ се използва общият индекс за коректно отразяване на състоянието във всички структури, за да не се показват като препоръки вече несъществуващи информационни ресурси. По този начин се поддържа актуалност и възпроизводимост на резултатите при нарастващ обем данни и натоварване.

Построените оперативни структури - разредената матрица „потребител-документ“, векторът на глобална популярност и матрицата на сходство между документите - осигуряват необходимата основа за генерирането на персонализирани препоръки. В следващия подраздел се използват именно тези оперативни структури, за да се формират персонализирани препоръки за всеки потребител при запазване на ниска латентност и възпроизводимост на резултатите.

Хибриден алгоритъм за генериране на „персонализирани препоръки“

Персонализираните препоръки се изчисляват върху предварително изградените структури: матрицата на сходство между документите S и разредената матрица „потребител-документ“ W . За всеки потребител u и кандидат-документ d оценката за релевантност се формира като сума от (1) съдържателен принос, пренесен от личната история на u чрез S , и (2) условен стабилизиращ индикатор за популярност, активен само при оскъдни данни за потребителя (хибридната оценката и функцията за резервната стратегия при „студен старт“ бяха формулирани в 3.6.2).

Хибридният алгоритъм, който се прилага, е следният:

1. формира се множество от кандидати, като се обединят k -най-близките съседи по матрицата на сходство S на всеки документ $i \in \text{история}(u)$
2. за всеки кандидат документ d се пресмята претеглената оценка спрямо историята на преглеждания и близките документи - оценка(u, d)
3. вече разглежданите от потребителя документи се изключват;
4. резултатите се подреждат по оценка и се връщат топ- k .

Гранични сценарии се обработват предвидимо:

- **Студен старт (нов/анонимен потребител).** При липса на история се ползват глобално популярните документи за формиране на начални предложения.

- **Изчерпана история.** Когато за профила няма нови съседи над прага по S, персоналният принос занулява изхода; системата преминава към резервна стратегия и предложения по популярност.
- **Смяна на интерес.** Единично ново взаимодействие извън установената тема действа като „котва“; неговите съседи според матрицата на сходство ускоряват пренасочването на препоръките към новия тематичен сегмент, без да се губи натрупаната история.
- **Смесен профил.** При разпределена история между няколко теми агрегиращата сума по историята на потребителя балансира приносите и поддържа разнообразие, без доминиране на един източник.
- **Студен старт за елементи.** Когато има документ, който е нов и няма история в регистрите на взаимодействия със системата, съдържателният компонент позволява релевантно извеждане спрямо вече познати на потребителя документи - проблем, който съвместното филтриране не решава.
- **Слабо представени поведенчески данни.** Ако историята на потребител е върху документи с ограничена аудитория и малко споделени читатели, съвместният показател е слаб, но съдържателният компонент през матрицата на сходство остава ефективен и осигурява смислово близки предложения.

Тази схема естествено извежда предимствата на хибридният алгоритъм. Съдържателният слой, реализиран чрез матрицата на сходство, гарантира смислова близост между документите, включително чрез локални съвпадения и тематични профили, и осигурява покритие за нови или слабо посещавани ресурси. Поведенческият слой, представен чрез матрицата „потребител-документ“, въвежда персонализация, основана на реалната история на взаимодействията. Индикаторът за глобална популярност се включва адаптивно като резервна стратегия при студен старт или оскъдна история и стабилизира резултатите, без да доминира, когато личната информация е достатъчна.

Разложимата форма на оценката запазва обяснимостта на препоръките чрез ясно разграничими приноси от историята и популярността и поддържа ниска латентност, тъй като работи изцяло върху предварително изчислени структури.

4.5.2. Експериментална валидация и тестови сценарии на модула за генериране на персонализирани препоръки

В тази подглава се проверява дали функционалният модул, който ползва хибридният алгоритъм за генериране на персонализирани препоръки (алгоритъм, базиран на елементно-ориентирано съвместно филтриране) проявява предвидимо и коректно поведение в ключови и гранични ситуации. Валидацията на системата е проведена чрез комбинация от тестове върху реални и синтетично генерирани данни, с цел да се оцени както поведението ѝ в практически условия, така и устойчивостта и релевантността ѝ при контролирани гранични сценарии. Реалните данни позволяват количествен анализ на работата на системата в контекста на действителни потребителски взаимодействия, докато синтетичният набор от данни е конструиран с цел по-прецизно и последователно тестване на специфични гранични случаи, които се срещат рядко или са недостатъчно представени в реалните данни. По този начин се осигурява както оценка на практическата приложимост на системата, така и възможност за контролирана проверка на нейното поведение при различни сценарии.

Първият етап от експерименталната оценка е насочен към анализ на поведението на системата в условия, максимално близки до реалната експлоатация. За тази цел е проведена количествена оценка върху реални регистри от потребителски взаимодействия, позволяваща да се проследи ефективността на предложения модел при естествени модели на използване и при реалистична разреденост на взаимодействията между потребителите и документите. Експерименталните тестове са проведени върху набор от данни, извлечени от платформата на дигиталната библиотека на Народна библиотека „Иван Вазов“ – Пловдив [166].

Количествената оценка е проведена върху данни, извлечени от регистрите на библиотечната платформа. Регистрите съдържат информация за достъп до документи, идентификатор на потребителя и времеви маркер на съответното взаимодействие. Всеки запис представлява имплицитно взаимодействие между потребител и документ и позволява проследяване на реалните модели на използване на библиотечните ресурси.

С цел осигуряване на контролируема и възпроизводима експериментална среда е използван представителен поднабор от библиотечната колекция. В предварителния етап на изследването са анализирани различни подмножества с размер между 200 и 1000 документа, което позволява да се оцени влиянието на размера на колекцията върху изчисляването на сходството между документите и върху изграждането на матрицата на

сходство. В окончателната конфигурация е избран набор от 500 документа, като регистрите са филтрирани така, че да включват единствено взаимодействия, свързани с тези документи. В рамките на настоящото изследване използването на контролиран поднабор позволява осигуряване на достатъчна плътност на взаимодействията и възпроизводимост на експериментите.

Използваният набор от данни съдържа 870 потребители, приблизително 12 000 взаимодействия и 500 документа. Както е характерно за дигиталните библиотечни среди, получената матрица потребител–документ е силно разрежена, тъй като всеки потребител взаимодейства само с ограничена част от наличните ресурси. Подобна структура на данните представлява съществено предизвикателство при изграждането на препоръчващи системи, тъй като затруднява откриването на устойчиви зависимости между потребителите и документите. Анонимните взаимодействия не са включени в количествената оценка, тъй като изграждането на потребителски профили и генерирането на персонализирани препоръки изискват наличие на идентифицируема история на взаимодействията.

За оценяване на качеството на препоръките е приложен подходът „оставяне на едно взаимодействие за тест“ (leave-one-out), широко използван при експериментални изследвания на препоръчващи системи. При всеки потребител последното регистрирано взаимодействие е отделено като тестов пример, а останалите взаимодействия са използвани за изграждане на потребителския профил. На тази основа системата генерира подреден списък с препоръчани документи, като ресурсите, с които потребителят вече е взаимодействал, се изключват от множеството на кандидатите.

В рамките на експерименталната оценка са сравнени три подхода за генериране на препоръки:

- Подход, базиран на елементно-ориентирано съвместно филтриране чрез използване на матрицата на взаимодействия потребител-документ;
- Подход, базиран на съдържателно филтриране чрез използване на семантична матрица на сходство между документите;
- Предложеният хибриден алгоритъм, комбиниращ поведенчески сигнали, семантично сходство и компонент, отчитащ глобалната популярност на ресурсите.

За всеки потребител е генериран списък от десет препоръчани документа, а качеството на препоръките е оценено чрез показателите Precision@10 [41], HR@10 [41] и NDCG@10 [202]. Показателят Precision@10 измерва дела на релевантните документи сред първите десет препоръки, HR@10 отчита дали релевантният документ присъства в препоръчания списък, а NDCG@10 оценява качеството на подреждането, като присъжда по-висока тежест на релевантните документи, разположени на по-предни позиции в списъка. Тези показатели са широко използвани при оценяването на препоръчващи системи, работещи с имплицитна обратна връзка [27], [41], [202].

Резултатите от количествената оценка са представени в Таблица 14. Наблюдава се, че предложеният хибриден алгоритъм постига най-високи стойности по всички използвани показатели - Precision@10, HR@10 и NDCG@10. Най-близки резултати до предложения хибриден алгоритъм показва подходът за съвместно филтриране, което е очаквано предвид факта, че оценката е извършена върху реални регистри за достъп и в значителна степен отразява съществуващите модели на колективно потребителско поведение. В тези условия поведенческите сигнали, извлечени от регистрите за достъп, съдържат достатъчно информация за изграждане на ефективни препоръки.

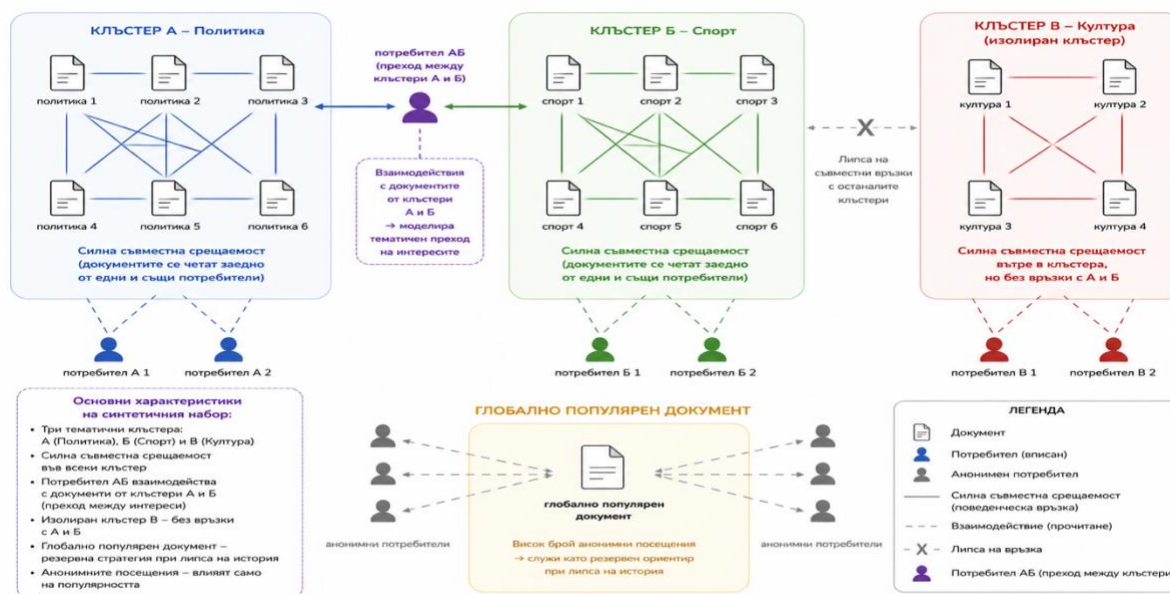
Таблица 14. Количествен анализ

<i>Подход</i>	<i>Precision@10</i>	<i>HR@10</i>	<i>NDCG@10</i>
<i>Съвместно филтриране</i>	0.0306	0.3057	0.1775
<i>Филтриране, базирано на съдържание</i>	0.0051	0.0506	0.0237
<i>Хибриден подход</i>	0.0321	0.3207	0.1888

Подобрието спрямо подхода, базиран единствено на елементно-ориентирано съвместно филтриране, е сравнително умерено. Въпреки това предложеният хибриден алгоритъм демонстрира по-стабилно поведение чрез комбиниране на поведенчески и семантични характеристики. Основното предимство на подхода се проявява при гранични случаи като ограничени взаимодействия, слабо свързани тематични области или нови потребители без история на взаимодействия, рязка промяна в потребителските интереси или появата на нови документи без натрупани взаимодействия. Обаче такива

сценарии се срещат сравнително рядко и не могат да бъдат систематично възпроизведени в реална среда. Поради тази причина е конструиран синтетичен набор от данни, позволяващ контролирано изследване на поведението на предложения алгоритъм при различни гранични случаи. Следва да се отбележи, че синтетичните сценарии представляват контролирана експериментална среда и не могат напълно да възпроизведат сложността и динамиката на реалното потребителско поведение. Въпреки това те позволяват систематично изследване на гранични случаи, които се срещат рядко в реалните регистри за достъп.

За да се минимизират случайните влияния и неравномерното разпределение на реално генерираните данни, синтетичният набор е организиран в три тематични групи документи - клъстери А, Б и изолиран В. Между документите в А и Б умишлено е въведена силна съвместна срещаемост, така че едни и същи потребители последователно да взаимодействат с документи от двете групи. По този начин се моделира наличие на ясно изразени поведенчески зависимости и сходни потребителски интереси. За разлика от тях, клъстер В е оставен без съвместни връзки с останалите документи, което позволява наблюдение на поведението на системата при липса на достатъчно съседни и ограничени поведенчески сигнали. Клъстерите и връзките между документите са показани на фигура 26.



Фигура 26. Структура на синтетичния набор от данни и взаимодействията между тематичните клъстери

Допълнително е включен документ с висок брой анонимни посещения, който изпълнява ролята на глобално популярен ресурс. Целта на този елемент е да се анализира поведението на системата в ситуации с ограничена или липсваща история на взаимодействия, при които компонентът за глобална популярност има по-значима роля при формирането на препоръките. По този начин синтетичният набор позволява отделните компоненти на хибридният алгоритъм да бъдат анализирани в контролирана среда и при предварително дефинирани условия.

Във всички експериментални сценарии е приложено ограничението документите, с които даден потребител вече е взаимодействал, да не бъдат предлагани повторно в препоръчания списък. Получените резултати са анализирани чрез съпоставяне между генерираните препоръки и очакваното поведение на системата за съответния сценарий. Този подход позволява по-прецизна оценка на устойчивостта на предложения алгоритъм и на способността му да генерира релевантни препоръки при наличие на ограничени, разредени или непълни данни.

За оценяване на поведението на предложения алгоритъм при различни условия на взаимодействие са дефинирани следните експериментални сценарии:

1) Студен старт при потребителя (липса на история / анонимен достъп)

Описание на сценария: Потребител без регистрирана история на взаимодействията или с анонимен достъп, при който векторът на потребителската активност е нулев.

Цел на теста: Да се валидира резервната стратегия на системата при липса на персонализираща информация.

Очакван резултат: Алгоритъмът следва да премине към неперсонализирана стратегия, основана на глобалната популярност, и да върне списък от k най-популярни ресурса. Препоръките трябва изрично да бъдат обозначени като базирани на популярност, а не на лична история.

2) Изчерпан изолиран тематичен клъстер

Описание на сценария: Потребител, който е консумирал изцяло съдържанието на затворен тематичен клъстер (напр. „Клъстер В“), без семантични връзки към други теми.

Цел на теста: Да се оцени поведението на системата при „семантична задънена улица“, когато всички релевантни ресурси вече са част от историята.

Очакван резултат: Системата не трябва да генерира персонализирани препоръки от изчерпания клъстер и следва да активира резервната стратегия, като предложи ресурси въз основа на глобална популярност.

3) Тематична преориентация (доминиращ интерес + единично ново взаимодействие)

Описание на сценария: Потребител с установена история в една доминираща тема (напр. „Клъстер А“), който осъществява единично взаимодействие с различна тема (напр. „Клъстер В“).

Цел на теста: Да се провери чувствителността на алгоритъма към нови интереси и способността му да реагира на тематични преходи.

Очакван резултат: Системата следва да разпознае новото взаимодействие като сигнал за промяна на интересите и да предложи ресурси от „Клъстер В“, дори при статистическата доминация на „Клъстер А“.

4) Смесен профил с балансиран интереси

Описание на сценария: Потребител с относително равномерно разпределена история между два независими тематични клъстера (А и В).

Цел на теста: Да се оцени способността на системата да комбинира множество интереси, без един от тях да доминира.

Очакван резултат: Препоръките трябва да включват балансиран набор от ресурси от двата клъстера, пропорционално на силата на поведенческите и съдържателните показатели, без „изтичане“ към нерелевантни или изолирани теми.

5) Студен старт при ресурса (нов документ)

Описание на сценария: Нов документ, който няма натрупани взаимодействия и няма общи читатели с други ресурси.

Цел на теста: Да се провери дали съдържателният компонент компенсира липсата на поведенчески данни за ресурса.

Очакван резултат: Документът следва да участва в персонализираните препоръки въз основа на съдържателна близост. За сравнение, при чисто съвместно филтриране, базирано на елементи, същият документ не би бил препоръчван.

6) Оскъдни и разредени поведенчески данни

Описание на сценария: Потребител с история, ограничена до ресурси с ограничен брой споделени взаимодействия.

Цел на теста: Да се оцени устойчивостта на алгоритъма при висока разреденост на поведенческите данни.

Очакван резултат: Съдържателният компонент следва да бъде водещ и да генерира смислово близки препоръки, компенсирайки слабите съвместни връзки. При класическо съвместно филтриране системата би преминала към резервна стратегия с популярни ресурси.

Оценяването обобщава два вида проверки: (1) качествена - дали списъците съвпадат с очакваната логика по сценарий (смяна на интерес, правилно преминаване към популярност и т.н.); и (2) количествена - дял на резултатите в първите k предложения и относителен принос на двата клъстера при смесен профил. Така резултатите от тестовете дават ясна картина, че модулът персонализира коректно, запазва полезност при липса на данни и реагира предвидимо при промяна на интересите. Резултатите от тестовете по сценариите са обобщени в Таблица 15.

Таблица 15. Сценарии и наблюдавани резултати

<i>Сценарий</i>	<i>Наблюдаван резултат</i>
<i>Студен старт при потребител</i>	Генерирани са популярни ресурси.
<i>Изчерпан изолиран тематичен клъстер</i>	Активирана е резервна стратегия - популярни ресурси.
<i>Тематична преориентация</i>	Препоръките се пренасочват към клъстер Б.

Смесен профил с балансиран интереси	Препоръките включват балансиран списък от двата кълстера на интереси на потребителя.
Студен старт при ресурс	Ресурсът участва в препоръките въпреки липсата на натрупани взаимодействия, благодарение на семантичната близост с вече посещавани ресурси.
Оскъдни и разредени поведенчески данни	Генерират се смислово близки препоръки, които компенсират слабите съвместни връзки.

Проведените експерименти показват, че предложеният алгоритъм демонстрира стабилно, предвидимо и устойчиво поведение както при реални потребителски взаимодействия, така и при контролирани гранични сценарии. Получените резултати потвърждават ефективността на хибридният подход при работа с разредени данни, ограничена история на взаимодействията и слабо свързани тематични кълстери, характерни за дигиталните библиотечни среди.

4.6. Обяснимост и етични механизми в реализацията

4.6.1. Обяснимост на алгоритмите

За да се повиши прозрачността на алгоритъма за извличане на „подобни документи“ и доверието в резултатите, числовите стойности на сходство не се показват директно, а се превеждат в разбираеми категории. Многокомпонентна оценка обединява различни показатели (семантично съдържание, тематична близост и налични именувани същности), поради което абсолютната ѝ стойност показва общата сила на връзката между два документа. В системата е заложена проста скала за интерпретация, която определя резултатите в следните нива:

- Идентичност ($S \geq 0.85$) - документите са почти еднакви (дубликати или препечатки).
- Високо сходство ($0.65 < S \leq 0.85$) - вероятност документите да споделят обща тема и конкретни общи обекти (например съвпадащи лица/организации).
- Свързани ($0.40 < S \leq 0.65$) - документите са в близки теми или подкорпуси, но може да описват различни събития или аспекти.

- Слаба връзка ($S \leq 0.40$) - слаби връзки; обичайно се филтрират в интерфейса, за да се ограничи информационният шум.

Тази интерпретация позволява „човешко четене“ на резултатите (т.е. кратко обяснение защо се предлага даден документ), без потребителят да трябва да познава използваните математически процедури. Праговете са конфигурируеми и могат да се прецизират емпирично според корпуса и наблюдаваното поведение.

4.6.2. Принципи за защита и минимизация на данните

При проектирането на системата са заложили механизми, гарантиращи сигурността на информацията и спазването на етичните норми за работа с данни (т.нар. „защита на личните данни още при проектирането“). Въпреки че обработваният корпус се състои предимно от публични архивни материали, архитектурата прилага следните мерки за ограничаване на риска:

- Минимизация на съхраняваните данни: В генерираните математически модели (вектори и матрици на сходство) не се пази оригиналният текст на документите. Използваните числови представяния са необратими, което означава, че от тях не може да бъде възстановено изходното съдържание. Това осигурява допълнително ниво на сигурност, тъй като самите изчислителни файлове са нечетими за човека.
- Логическо разделение: Информацията за извлечените именуванни същности (лица и организации) се съхранява в отделни структури, а връзката с документите се осъществява единствено чрез служебни идентификатори, което предотвратява прякото асоцииране без специализиран достъп.
- Защита на потребителската активност: Регистрите на достъп се ползват от препоръчващия модул единствено чрез псевдонимизирани потребителски идентификатори, без да се съхранява достъп до профилни данни или история на търсенето, която би могла да послужи за профилиране.
- Анонимизиран анализ: В етапа на тестване и настройка на системата, диагностичните инструменти оперират само с обобщена статистика (напр. процент на съвпадение), без да записват или излагат конкретни текстови откъси в системните регистри.

4.7. Ограничения и валидност на предложената архитектура

Предложената архитектура е съобразена със спецификата на наличния корпус, типовете регистри на взаимодействията и избраните методи за обработка на текст. За да бъде оценена коректно тяхната приложимост, е необходимо ясно да се очертаят основните ограничения и обхватът на валидност на резултатите. Този раздел систематизира допусканията, техническите изисквания и факторите, които могат да повлияят върху качеството и устойчивостта на препоръките.

4.7.1. Ограничения, свързани с данните

1. Специфика на корпуса

Корпусът включва предимно многотематични периодични издания на български език. Получените модели и наблюдения са най-надеждни в рамките на този контекст. При пренасяне към други езици или силно различни тематични области е необходима повторна калибрация и допълнителна оценка.

2. Качество на текстовото съдържание

Част от текстовете могат да съдържат грешки от оптично разпознаване (OCR), липсващи сегменти или неструктурирани елементи. Тези неточности пряко влияят върху векторните представяния, матрицата на сходство и извлечените именуван същности и следователно представляват източник на шум, който не може да бъде напълно елиминиран.

4.7.2. Ограничения на съдържателния модел

1. Зависимост от предварително обучени модели

Векторните представяния на текстовете се базират на предварително обучени езикови модели. Те не са оптимизирани специално за всеки възможен домейн и могат да не улавят в пълна степен специфични термини, остарели езикови форми или редки имена. Това ограничава точността на измерената семантична близост.

2. Качество на именуваните обекти

Използването на именуван същности като допълнителен индикатор зависи от надеждността на модула за разпознаване. Грешно разпознати или пропуснати обекти могат да доведат до неправилно приближаване или отдалечаване на документи.

Филтрирането по честота намалява влиянието на тези грешки, но не ги отстранява напълно.

3. Избор на параметри

Комбинирането на различните компоненти на сходството и прилагането на размито групиране изискват избор на прагове и коефициенти. Тези параметри влияят върху структурата на матрицата на сходство и се определят емпирично, което трябва да се отчита при тълкуване на резултатите.

4.7.3. Ограничения на поведенческите данни

1. Имплицитен характер на показателите

Регистрираните прегледи и действия се третираат като положителни индикатори, без изрични оценки от потребителите. Прегледът обаче не винаги означава одобрение или удовлетворение, също не означава, че въпросният текст е бил прочетен, а само че е бил отворен. Това води до неизбежна несигурност, която се компенсира частично чрез използване на множество източници (съдържание, популярност), но не може да бъде напълно премахната.

2. Неравномерно разпределение

Някои документи събират значителен брой взаимодействия, докато други остават почти без история. Хибридният алгоритъм смекчава този ефект, като използва матрицата на сходство и съдържателната близост, но при документи с много ограничени данни точността на препоръките остава по-ниска.

3. Идентичност на потребителите

Анонимни потребители са агрегирани заедно, не е правен опит за идентифициране на отделни сесии или профили.

4.7.4. Технически и изчислителни ограничения

1. Изчислителна тежест

Изграждането на векторните представяния, матрицата на сходство и тематичните модели изисква значителни изчислителни ресурси (процесорно време, памет, при възможност графични ускорители). Това налага:

- разделяне на обработката на отделни стъпки и пакетен режим;

- ограничаване на съхраняваните стойности на сходство до най-значимите;
- внимателен избор на периодичност за пълно обновяване, ако изобщо се позволява пълно обновяване.

2. Големи езикови модели за именувани същности

При използване на големи езикови модели за извличане и проверка на именувани същности се увеличава изчислителната цена и се поставят допълнителни изисквания към хардуера. В практическа среда това може да изисква специализирани сървъри или комбинация от локална и отдалечена обработка. Тези особености трябва да се вземат предвид при пренасяне на предложеното решение в други или по-големи среди.

3. Пакетно обновяване

Поради избрания подход за пакетна обработка препоръките не отразяват всяко ново взаимодействие в реално време. Между две обновявания съществува период, в който моделът работи с леко остарели данни. Това е компромис между точност и изчислителна приложимост.

4.7.5. Валидност на резултатите

- Вътрешна валидност

Процесите по подготовка на данните, изграждане на представянията и конструиране на оперативните структури са ясно дефинирани и повторяеми. Това позволява възпроизвеждане на експериментите при същите настройки и дава основания да се счита, че наблюдаваните разлики между конфигурациите се дължат на използваните методи, а не на случайни фактори.

- Външна валидност

Моделът е директно приложим за среди с подобна структура на съдържанието (документни колекции, дигитални библиотеки) и сходен начин на регистриране на взаимодействията. При прилагане в различен контекст е необходимо адаптиране на параметрите, повторно обучение или подмяна на отделни компоненти (например езиковия модел), както и нова оценка на качеството.

Този раздел не цели да омаловажи предложената архитектура, а да очертае ясно границите, в рамките на които тя е коректна и полезна. Осъзнаването на тези

ограничения е предпоставка за реалистична интерпретация на резултатите и за планиране на бъдещи подобрения.

4.8. Обобщение

В тази глава е представена реализацията на предложената архитектура за персонализирано представяне на съдържание в дигитална библиотека, изградена върху ясно разграничение между асинхронен слой за изчислително тежка предварителна подготовка и лек интерактивен слой, работещ с предварително изчислени структури. Реализацията включва три основни компонента в асинхронния слой: (1) услуга за извличане на именувани същности, използвана за семантично обогатяване на метаданните; (2) модул за изграждане на матрица на сходство между документите чрез векторни представяния и многокомпонентна оценка на близост; и (3) модул за обработка на регистрите за потребителски взаимодействия, чрез който се конструира разредената матрица „потребител-документ“ и се изчислява индикатор за глобална популярност.

Услугата за именувани същности функционира като самостоятелен модул за извличане на структурирани фактологични показатели, които се използват като допълнителен семантичен слой при оценката на сходството. В разделите, посветени на матрицата на сходство, са описани процесите по синонимна нормализация, извличане на векторни представяния и формулиране на многокомпонентна функция за документно-документно сходство, интегрираща семантични, тематични и фактологични характеристики. Модулът за персонализирани препоръки реализира агрегиране на имплицитната обратна връзка в матрицата „потребител-документ“ и адаптивно включване на глобалната популярност при оскъдни данни.

Експерименталната валидация разглежда ефективността на всеки от основните компоненти. Услугата за именувани същности е валидирана чрез качествена проверка на коректността и чрез измерване на производителността върху представителна извадка от документи. Модулът за „подобни документи“ е подложен на параметрично калибриране и структурен анализ на матрицата на сходство, като резултатите показват, че включването на тематични профили и именувани същности води до по-прецизно моделиране на близостта в многотематичен корпус, въпреки умерено увеличената изчислителна сложност. Персонализираните препоръки са изследвани чрез синтетични сценарии, обхващащи студен старт, смяна на интереси, изчерпване на тематични

кълстери, смесени профили и разредени поведенчески данни, като наблюденията потвърждават стабилно, обяснимо поведение и ниска латентност.

Наред с това са разгледани механизми за обяснимост, основани на разложима оценка и интерпретируема скала за релевантност, както и принципи за защита и минимизация на данните. Описани са и процедурите за инкрементално обновяване, паралелна обработка и използване на компактни структури, които осигуряват мащабируемост и устойчивост при нарастващ обем съдържание и натоварване.

В заключение, реализираната система потвърждава практическата приложимост на предложената архитектура. Многокомпонентното съчетаване на съдържателна близост, поведенчески показатели и обогатени метаданни води до ефективни и обясними персонализирани резултати. Експерименталните резултати подкрепят формулираните в предходната глава хипотези, че (1) многокомпонентната функция за документно-документно сходство по-добре улавя структурата на корпуса и (2) персонализацията, комбинираща история, съдържателен контекст и популярност, осигурява по-устойчиви и качествени препоръки в гранични сценарии. Това създава стабилна основа за последващо разширяване към допълнителни корпуси, езици и интеграция с оперативни библиотечни системи.

ГЛАВА 5. ЗАКЛЮЧЕНИЕ – РЕЗЮМЕ НА ПОЛУЧЕНИТЕ РЕЗУЛТАТИ

5.1. Обобщение на резултатите

В дисертационния труд бе разработена и валидирана цялостна архитектура за персонализирано представяне на съдържание в дигитални библиотеки, която интегрира три допълващи се източника на информация: (а) съдържателни представяния на документите и многокомпонентна оценка на близост (семантика, локални съвпадения, тематична компонента и компонента от именувани същности); (б) поведенчески данни под формата на разрежена матрица „потребител-документ“ с имплицитни тежести и показатели за глобална популярност; (в) обогатени метаданни от услуга за извличане на именувани същности, пригодена за български език и за корпуси с дълги текстове и нееднородно качество.

Архитектурата е реализирана модулно, с ясно разграничение между изчислително интензивните процеси (подготовка на оперативни структури), изнесени в асинхронен слой, и лек интерактивен слой за бързо обслужване на заявки. Това осигурява ниска латентност, възпроизводимост и мащабируемост. Експерименталната верификация - върху реални и синтетични данни - показва устойчиво и предвидимо поведение на двата ключови изхода: „подобни документи“ (елементно ориентирана навигация) и „персонализирани препоръки“ (агрегиране на съдържателна близост през личната история с адаптивен принос на популярността). Включването на именувани същности повишава обяснимостта и стабилизира оценката при шум и дълги текстове, а тематичният слой намалява фалшивите сходства при обща, но неинформативна лексика.

Изследването завършва с работеща архитектура и прототипи, които демонстрират, че интегрирането на съдържателни представяния, поведенчески показатели и метаданни води до по-уместни, обясними и мащабируеми препоръки в дигитална библиотека. Доказателството е емпирично: калибрирана и валидирана матрица на сходство; разрежена матрица „потребител-документ“ с функция на насищане; хибридна функция за персонализация с адаптивна популярност; и последователни резултати в контролирани сценарии („студен старт“, изчерпана история, тематичен завой, смесен профил), където хибридният подход превъзхожда единични (само съдържателни или само поведенчески) решения.

Поетапно бяха изпълнени задачите:

1. Извършен е аналитичен литературен преглед с идентифициране на актуалните тенденции и проблеми (оскъдни поведенчески данни, студен старт, обяснимост, мащабируемост);
2. Предложени са концептуален модел и архитектурна рамка с ясно разграничение между предварителната обработка на данните, включително изграждане на оперативни структури, и интерактивната част, която реализира крайния етап на генериране на двата типа препоръки;
3. Реализирана е услуга за извличане на именувани същности и два основни модула - за генериране на матрица на сходство и за матрица „потребител-документ“/популярност - със съгласувани идентификатори;
4. Проведена е експериментална валидация върху синтетични данни и реална извадка от дигитална библиотека, потвърждаваща очакваното поведение на препоръчителния алгоритъм и навигацията по сходство;
5. Дефинирани са механизми за обяснимост (разлагане на оценката по източници на показатели и лингвистични етикети на интервалите на сходство) и принципи за защита и минимизация на данните.

5.2. Насоки за бъдещо развитие

Перспективите за развитие са в няколко направления.

1. **Разширяване на показателите:** времеви и контекстни фактори (време на четене на документ, до къде е стигнал потребителят в документ - ако документът е само отворен, може да говорим за имплицитна отрицателна оценка, а не за положителна и т.н.), по-богати метаданни (жанрове, таксономии - поради липсата на тематика на текстовете не може да се предлага документ на потребител, защото той се интересува от тази тематика, а се разчита само на подобност на текстовете), многоезични корпуси.
2. **Динамично дообучаване и персонализирано калибриране:** инкрементални алгоритми за адаптация на теглата и праговете според динамиката на колекцията и променящите се интереси; самообучаващи се елементи и обучение с подкрепление при достатъчен мащаб.

3. **Инфраструктурна оптимизация:** приблизително търсене на най-близки съседи (приблизителни индекси за NN), по-ефективни схеми за съхранение (орязани графи).
4. **Реално оценяване с участието на реални потребители:** A/B експерименти, регистри на взаимодействие и качествени изследвания за възприемана релевантност, доверие и полезност на обясненията.

Предложената архитектура показва, че съгласуваното интегриране на съдържателни, поведенчески и семантични показатели води до ефективна, обяснима и мащабируема персонализация в дигитални библиотеки. Многокомпонентна оценка за документно-документно сходство подобрява улавянето на тематичната структура в многотематични корпуси, а хибридната персонализация повишава качеството на препоръките при оскъдни или изкривени поведенчески данни.

Естественото продължение на изследването е пренасянето към продукционна среда с реални потребители, разширяване на набора от показатели и въвеждане на механизми за динамична адаптация, с цел емпирично валидиране на практическата приложимост и устойчивостта на предложеното решение в дългосрочна експлоатация.

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.

СПИСЪК НА АВТОРСКИТЕ 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 НА 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)*, ACM International Conference Proceedings Series, ACM, New York, USA, 2024, ISBN:979-8-4007-1684-3/24/06, DOI:10.1145/3674912.3674935, 227-231. SJR (Scopus) : 0.253

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

СПИСЪК НА 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: **Е. 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: **Е. Mitreva**, V. Georgiev, A. Nikolova, , Дата: 15.06.2023
3. Изнесен научен доклад на международната UNESCO конференция Digital Presentation and Preservation of Cultural and Scientific Heritage (DiPP2023), 07-10 Септември 2023, Бургас, България, Тема: Personalization Approaches for Cultural Heritage Study, Автори: **Е. 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 НА ФИГУРИТЕ

Фигура 1. Етапи на подхода анализ на потребителското поведение в уеб среда.....	21
Фигура 2. Данни за класификация от Машина на опорните вектори	26
Фигура 3. Два класификатора от Машина на опорните вектори	26
Фигура 4. Хиперравнина в Машина на опорните вектори	27
Фигура 5. По-сложни данни за разделяне от Машина на опорните вектори	27
Фигура 6. Алгоритъм на к-средни (k-means)	29
Фигура 7. Използване на CountVectorizer	36
Фигура 8. Честота на използваните техники за персонализация в анализираните публикации	50
Фигура 9. Компонентна диаграма на концептуалния модел	63
Фигура 10. Асинхронен слой.....	65
Фигура 11. Диаграма на случаи на употреба за „подобни документи“	66
Фигура 12. Диаграма на случаи на употреба за „Персонализирани препоръки“	67
Фигура 13. Подготовка на текстовите обекти.....	72
Фигура 14. Диаграма на последователността на създаване на оперативните структури.....	82
Фигура 15. Концептуална схема на персонализирани препоръки	84
Фигура 16. Архитектура на модулите за предоставяне на персонализирано съдържание	94
Фигура 17. Блок диаграма на услугата за извличане на същности.....	97
Фигура 18. Пример за формат на per.json.....	100
Фигура 19. Синонимен речник	100
Фигура 20. Процес на генериране на матрица на сходство	102
Фигура 21. Тематични клъстери и връзки между документите в синтетичния набор от данни	114
Фигура 22. Топлинна карта на сходството между документите при използване единствено на косинусова близост	115
Фигура 23. Топлинна карта на сходството между документите при многокомпонентна оценка	116
Фигура 24. Практическа демонстрация на „подобни документи“	118
Фигура 25. Генериране на разрежена матрица „потребител-документ“	120
Фигура 26. Структура на синтетичния набор от данни и взаимодействията между тематичните клъстери	128

СПИСЪК НА ТАБЛИЦИТЕ

Таблица 1. Сравнителен анализ между традиционни и дигитални библиотеки	14
Таблица 2. Сравнение между подходите и методите за персонализация	33
Таблица 3. Сравнение на основните методи на кодиране на текст	39
Таблица 4. Подходи за персонализацията в дигиталните библиотеки: силни и слаби страни	44
Таблица 5. Честота на използваните техники за персонализация в анализираните публикации	48
Таблица 6. Основни области на ограничения в литературата.....	51
Таблица 7. Пропуски и насоки за бъдещи изследвания и подобрения	53
Таблица 8. Сравнение на техники за намаляване на размерността при дълги текстове	74
Таблица 9. Изчислителни характеристики и времена за изпълнение на моделите	98
Таблица 10. Времена за обработка при различни размери на сегмента.....	99
Таблица 11. Роля на компонентите в многокомпонентна крайна оценка	105
Таблица 12. Аблационен анализ.....	112
Таблица 13. Примерни препоръки, съответстващи на експерименталния набор	117
Таблица 14. Количествен анализ.....	127
Таблица 15. Сценарии и наблюдавани резултати.....	131

РЕЧНИК НА ИЗПОЛЗВАНИТЕ ТЕРМИНИ И СЪКРАЩЕНИЯ

Аблационен анализ (Ablation Study) - Систематичен експеримент, при който се изключват или заменят отделни компоненти на модел (или входни показатели), за да се измери техният самостоятелен принос към крайния резултат. Сравняват се метрики на „пълната“ конфигурация срещу редуцирани варианти, за да се установи кои елементи са необходими/достатъчни и колко допринасят.

Анализ на главните компоненти (Principal Component Analysis - PCA) - статистически метод за редуциране на размерността, който чрез ортогонална трансформация извлича некорелирани главни компоненти, обобщаващи максималната вариация в данните.

Анализ на потребителското поведение в уеб среда (Web Usage Mining) - Процес на извличане на модели и знания от данни, описващи поведението на потребителите в уеб среда - включително анализ на дневници на уеб сървъри, последователности на кликания и взаимодействия с уебсайтове.

Векторизация на данни (Vectorizing data) - Процес на преобразуване на данни, представени в нематематическа форма (например текст, изображения или категории), в числови вектори, подходящи за обработка от алгоритми за машинно обучение. Векторизацията позволява на моделите да интерпретират и обработват съдържанието чрез математически операции.

Големи данни (Big Data) - термин, използван за обозначаване на изключително големи и разнообразни обеми от данни, които се генерират с висока скорост и не могат да бъдат обработени ефективно с традиционни методи за управление на данни.

Големи езикови модели (Large Language Models, LLMs) - Невронни модели, обучени върху огромни количества текстови данни, които могат да разбират, обобщават и генерират човешки език, използвайки техники от дълбокото обучение.

Ефект на филтърния балон (Filter Bubble Effect) - явление, при което алгоритмите за персонализация на съдържанието в интернет (напр. в социални мрежи или търсачки) показват на потребителя предимно информация, съответстваща на неговите предишни интереси и възгледи. Това ограничава достъпа до разнообразни гледни точки и може да доведе до информационна изолация и поляризация на мненията.

Извличане на характеристики (Feature extraction) - Процесът на идентифициране и извличане на значими признаци или атрибути от сурови данни, които съдържат полезна информация за решаване на конкретна задача (напр. класификация или регресия). Целта е да се намали размерността и да се улесни последващият анализ или обучение на модел.

Изкуствен интелект (Artificial Intelligence) - Област от информатиката, която разработва методи и системи, способни да изпълняват задачи, изискващи човешка интелигентност - като възприятие, разсъждение, учене и вземане на решения.

Именувана същност (Named Entity) - Обект от реалния свят, който може да бъде еднозначно идентифициран с име - например човек, организация, място, дата или продукт.

Инженеринг на характеристики (Feature Engineering) - Създаване, модифициране или комбиниране на характеристики (признаци) с цел подобряване на представянето на модела. Включва техники като мащабиране, нормализация, трансформации и създаване на нови променливи от наличните данни.

Компресиран редови разреден формат (Compressed Sparse Row, CSR) - Представяне на разредена матрица, при което се съхраняват само ненулевите стойности, индексите на колоните им и указатели към началото на всеки ред. Подходящ за бързи редови операции и матрично-векторни умножения, с малък паметен отпечатък при силно разредени данни.

Лематизация (Lemmatization) - Процес на преобразуване на думите до тяхната основна, речникова форма - лема. Например думите „ходя“, „ходеше“, „ходили“ се свеждат до общата им лема „ходя“. Лематизацията отчита граматичната и морфологичната информация на думата, което я прави по-точна, но и по-ресурсоемка техника за нормализация на текста.

Машинно обучение (Machine Learning) - Подобласт на изкуствения интелект, която изучава алгоритми и методи, позволяващи на компютърни системи да се обучават от данни и да подобряват представянето си без изрично програмиране.

Метод на k-най-близките съседи (K-Nearest Neighbors, KNN) - метод в машинното обучение, използван за класификация и регресия. Алгоритъмът определя класа или стойността на дадена наблюдавана точка въз основа на мнозинството (или

средната стойност) от нейните k най-близки съседи в обучаващия набор според избрана метрика за разстояние (напр. Евклидово или Манхатънско разстояние).

Метод на k -средните (k-means) - Алгоритъм за клъстеризация, който разделя набор от данни на предварително определен брой клъстери (k), като минимизира разстоянието между отделните точки и центроида на съответния клъстер. Работи чрез итеративно пренасочване на наблюденията към най-близкия центроид и актуализиране на позициите на центроидите до постигане на стабилно разпределение.

Метод на размити k -средни (Fuzzy k-means) - Вариант на алгоритъма k-means, при който всяка наблюдавана точка може да принадлежи частично към повече от един клъстер с определена степен на принадлежност (стойност между 0 и 1). Този подход позволява по-гъвкаво моделиране на данни с припокриващи се структури и се използва в области като анализ на модели и разпознаване на образи.

Метод на опорните вектори (Support Vector Machine, SVM) - Алгоритъм за машинно обучение, използван основно за класификация и регресия, който намира оптимална граница (хиперплоскост), разделяща различни класове данни с максимален марж.

Обработка на естествен език (Natural Language Processing, NLP) - Област от изкуствения интелект и компютърната лингвистика, която се занимава с разработването на методи и алгоритми за автоматичното разбиране, анализ и генериране на човешки език от компютърни системи.

Обучение с частичен надзор (Semi-supervised Learning, SSL) - Подход в машинното обучение, който комбинира малко количество етикетирани данни с голямо количество неклассифицирани данни, за да подобри точността на модела. Използва се, когато етиктирането на данните е скъпо или трудоемко, а неклассифицирани данни са лесно достъпни.

Подсилващо обучение (Reinforcement Learning) - Метод на машинно обучение, при който агентът взаимодейства със среда, като извършва действия и получава обратна връзка под формата на награди или наказания. Целта му е да се научи на оптимална стратегия (policy), която максимизира натрупаната награда във времето.

Пренапасване (Overfitting) - Състояние, при което моделът в машинното обучение се приспособява прекалено към обучаващите данни и губи способност да обобщава върху нови примери.

Проблем на студения старт (Cold Start Problem) - Проблем в системите за машинно обучение и препоръчване, който възниква, когато липсват достатъчно данни за нов потребител, елемент или услуга, което затруднява генерирането на точни препоръки или моделиране на поведение.

Проклятие на размерността (Curse of Dimensionality) - Явление в статистиката и машинното обучение, при което увеличаването на броя на измеренията (променливите) в пространството води до експоненциално нарастване на необходимото количество данни и изчислителни ресурси. При висока размерност точките стават разпределени „рядко“ в пространството, което намалява ефективността на методи, зависещи от разстояния или близост (като k-най-близките съседи).

Равномерна апроксимация и проекция на многообразия (Uniform Manifold Approximation and Projection, UMAP) - нелинеен метод за редуциране на размерността, който моделира данните като многомерно многообразие и извършва проекция в ниско измерение при максимално запазване на локалната и глобалната структура.

Разпознаване на именувани единици (Named-Entity Recognition, NER) - Задача в обработката на естествен език, която автоматично открива и класифицира именувани единици в текст според техния тип (например личности, организации или местоположения).

Регистри (за достъп) (Logs) - Файлове или бази от данни, в които систематично се записват събития, действия или транзакции, извършвани в рамките на информационна система, с цел последващ анализ, мониторинг и проследимост.

Решетъчно търсене на хиперпараметри (Grid Search) - Метод за избор на хиперпараметри чрез предварително дефинирана дискретна решетка от възможни стойности и систематично оценяване на всяка комбинация според избрана метрика. Цел: да се намери конфигурация с най-добро емпирично поведение.

Сдвиг на домейна (Domain Shift) при PCA - Промяна в разпределението на новите данни спрямо това, върху което е обучена PCA, при което главните компоненти престават да описват релевантната вариация. Резултат: изкривени проекции и деградация на метриците, освен ако проекцията не се преобучи/адаптира.

Система за препоръки (Recommendation System) - Информационна система, която анализира поведението, предпочитанията и миналите действия на потребителите, за да предлага персонализирани продукти, услуги или съдържание.

Стеминг (Stemming) - Процес на свеждане на думите до техния корен (stem) чрез премахване на наставки, представки или окончания, без задължително да се получава граматически правилна форма. Например „говоря“, „говореше“ и „говорим“ могат да се сведат до „говор“. Стемингът е по-бърз, но по-малко точен от лематизацията.

Стоп думи (Stop Words) - Думи, които се срещат често в езика, но обикновено не носят съществена смислова стойност при автоматична обработка на текст (например „и“, „в“, „на“, „че“). При предварителната обработка на текстови данни тези думи често се премахват, за да се намали шумът и да се подобри ефективността на алгоритмите за анализ на естествен език.

Съвместно филтриране (Collaborative Filtering) - Метод за изграждане на системи за препоръки, който анализира сходствата между потребители или между техните взаимодействия с обекти (например оценки, кликания или прегледи), за да предлага нови елементи, основани на поведението на сходни потребители.

Факторизация на неотрицателни матрици (Non-negative Matrix Factorization, NMF) - метод за редуциране на размерността, който разлага дадена неотрицателна матрица на произведение от две по-малки неотрицателни матрици, като по този начин извлича интерпретируеми адитивни компоненти в данните.

Филтриране, базирано на елементи (Item-based Filtering) - Вариант на съвместно филтриране, при който препоръките се генерират чрез изчисляване на сходство между самите елементи, а не между потребителите. Често се използва за мащабируеми системи, тъй като сходствата между елементи се променят по-бавно от тези между потребители.

Филтриране, базирано на съдържание (Content-based Filtering) - Метод за препоръчване, който анализира характеристиките на елементите (например описание, ключови думи, категории) и предлага нови обекти, сходни по съдържание на вече харесаните или използвани от конкретния потребител.

Функция на насищане (Saturation Function) - Кратка нелинейна трансформация на суров показател (напр. брой прегледи), при която прирастът на стойността намалява с увеличаване на входа и се ограничава с горна граница. Цел: да се редуцира влиянието на многократни/шумни действия и да се стабилизират имплицитните оценки.

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] M. Fekadu и D. Alemneh, „Digital Library Models: A Systematic Review,“ в *In Sustainability and Empowerment in the Context of Digital Libraries: 26th International Conference on Asia-Pacific Digital Libraries*, Singapore, 2024. ISBN 978-981-19-0351-9, https://doi.org/10.1007/978-981-96-0865-2_7.
- [4] C. R. A. Owusu-Ansah, „Digital Information and Library Services in ODDE,“ в *Handbook of Open, Distance and Digital Education*, 2022. https://doi.org/10.1007/978-981-19-0351-9_45-1.
- [5] 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>.
- [6] R. Prasanna и S. Yogendra, „The Role of Artificial Intelligence in Enhancing Digital Library Services,“ *International Journal of Emerging Technologies and Innovative Research*, том 10, № 12, pp. 712-723, 2023.
- [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. ISSN: 2340-1095, doi: 10.21125/iceri.2017.0595.
- [8] D. Paneva-Marinova, „A Semantic-Oriented Architecture of a Functional Module for Personalized and Adaptive Access to the Knowledge in a Multimedia Digital Library,“ *International Journal „Serdica Journal of Computing”*, том 2, № 4, pp. 403-424, 2008. doi: 10.55630/sjc.2008.2.403-424.
- [9] M. Goynov, D. Luchev, D. Paneva-Marinova, R. Pavlov и K. Rangochev, „Full-fledged Access and Usability of Content in Digital Cultural Heritage Library: Approaches, Paradigms and Implementation,“ *ACM Journal on Computing and Cultural Heritage*, том 17, № 1, pp. 1-12, 2024. doi: <https://doi.org/10.1145/3631135>.
- [10] 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.
- [11] M. Goynov, D. Luchev, D. Paneva-Marinova, S. Najdenova, L. Zlatkov, L. Pavlova и E. Pilege, „Digital Revival of the Bulgargica Collection of the Central Library of the Bulgarian Academy of Sciences,“ *Digital Presentation and Preservation of Cultural and Scientific Heritage*, том 13, pp. 77-86, 2023. <https://doi.org/10.55630/dipp.2023.13.7>.
- [12] D. Paneva-Marinova, M. Goynov и D. Luchev, „Multimedia Digital Library as a Constructive Block in Ecosystems for Digital Cultural Assets,“ *Digital Presentation and Preservation of Cultural and Scientific Heritage*, том 7, pp. 31-40, 2017. doi: 10.55630/dipp.2017.7.2.
- [13] L. Pavlova-Draganova, D. Paneva-Marinova, R. Pavlov и M. Goynov, „On the Wider Accessibility of the Valuable Phenomena of Orthodox Iconography through Digital Library,“ в *Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010)*, 2010.
- [14] 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>.
- [15] 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.

- [16] 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.
- [17] G. Di Nunzio, „Focused Issue on Digital Library Challenges to Support the Open Science Process,” *International Journal on Digital Libraries*, том 24, № 4, p. 185–242, 2023. doi: <https://doi.org/10.1007/s00799-023-00388-9>.
- [18] S. Mukherjee и S. K. Patra, „Digital Library Initiatives in India: A Comprehensive Study,” *arXiv preprint arXiv:2303.13594*, 2023. doi: 10.48550/arXiv.2303.13594.
- [19] 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.
- [20] B. J. Bamgbade, B. A. Akintola, D. O. Agbenu, C. O. Ayeni, F. O. O. и H. O. & Abubakar, „Comparative analysis and benefits of digital library over traditional library,” *World Scientific News*, 2015.
- [21] 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>.
- [22] 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.
- [23] S. K. Bhat , „The Role of Digital Libraries in Enhancing Information Accessibility: A Study,” *International Journal for Multidisciplinary Research (IJFMR)*, том 6, № 6, 2024. E-ISSN: 2582-2160, doi: <https://doi.org/10.36948/ijfmr.2024.v06i06.32585>.
- [24] D. Hapsari, H. Haryanto и A. G. Firdausy, „Digital library innovation and challenges in supporting sustainable development through digital transformation,” *BIS Information Technology and Computer Science*, том 2, 2025. doi: 10.31603/bistycs.186.
- [25] 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.
- [26] 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.
- [27] Z. Fayyaz, M. Ebrahimian, 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>.
- [28] 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>.
- [29] 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.
- [30] E. Mitreva, A. Nikolova and V. Georgiev, „Web Mining Techniques Applicable for Cultural Heritage Observations,” *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 11, pp. 253-260, 2021. doi: <https://doi.org/10.55630/dipp.2021.11.22>.
- [31] J. Stoikov, „Using Conditional Probability for Discovering Semantic Relationships between Named Entities in Cultural Heritage Data,” *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 11, pp. 77-88, 2021. <https://doi.org/10.55630/dipp.2021.11.7>.

- [32] Y. Deng, Y. Li, W. Zhang, B. Ding and W. Lam, "Toward Personalized Answer Generation in E-Commerce via Multi-perspective Preference Modeling.," *ACM Transactions on Information Systems*, vol. 40, no. 4, pp. 1-28, 2022. doi: <https://doi.org/10.1145/35077>.
- [33] L. Narke and A. Nasreen, "A comprehensive review of approaches and challenges of a recommendation system," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 4, pp. 381-384, 2020. ISSN: 2581-5792 .
- [34] H. Ju and H. Wang, "Application analysis of computer web data mining technology in E-commerce.," in *5th International Conference on Electronic Information Technology and Computer Engineering*, 2021. doi: <https://doi.org/10.1145/3501409.350162>.
- [35] J. D. T. Nugroho, R. Mahendra and I. Budi, "Web Mining in e-Procurement: A Case Study in Indonesia," in *In Proceedings of the 2021 3rd Asia Pacific Information Technology Conference (APIT '21)*, New York, 2021. doi: <https://doi.org/10.1145/3449365.3449382>.
- [36] M. Srivastava, R. Garg and P. K. Mishra, "Analysis of data extraction and data cleaning in web usage mining," in *International Conference on Advanced Research in Computer Science Engineering & Technology*, 2015. doi: <https://doi.org/10.1145/2743065.2743078>.
- [37] M. Manchanda, "Web Usage Mining: Dynamic Methodology to Preprocessing Web Logs," *Helix*, vol. 8, no. 5, pp. 3810-3815, 2018. doi: 10.29042/2018-3810-3815.
- [38] S. Marsden, "Search Engine Crawling," 17 05 2018. [Online]. Available: <https://www.lumar.io/learn/seo/crawlability/search-engine-crawling/>. [Accessed 13 10 2023].
- [39] M. Liao, S. S. Sundar and J. Walther, "User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering," in *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2021. doi: <https://doi.org/10.1145/3491102.3501936>.
- [40] 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.
- [41] 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.
- [42] 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>.
- [43] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu and L. He, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1-41, 2022. doi: <https://doi.org/10.1145/3495162>.
- [44] J. Chen, Z. Gong and W. Liu, "A Dirichlet process biterm-based mixture model for short text stream clustering," *Applied Intelligence*, vol. 50, no. 5, pp. 1609-1619, 2020. doi: 10.1007/s10489-019-01606-1.
- [45] 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.
- [46] 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.
- [47] X. Zhang, F. Guo, C. Tao , P. Lei, G. Beliaikov 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>.

- [48] 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.
- [49] T. Alasali и Y. Ortakci, „Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications,“ *Anatolian Science*, том 9, № 1, pp. 32-50, 2024. doi: 10.53070/bbd.1421527.
- [50] S. Naeem, A. Ali, S. Anam и M. M. Ahmed, „An Unsupervised Machine Learning Algorithms: Comprehensive Review,“ *International Journal of Computing and Digital Systems*, том 13, № 1, pp. 911 - 921, 2023. doi: 10.12785/ijcds/130172.
- [51] A. E. Mehyadin и A. M. Abdulazeez, „Classification based on semi-supervised learning: a review,“ *Iraqi Journal for Computers and Informatics*, том 47, № 1, 2021. doi: 10.25195/ijci.v47i1.277.
- [52] Y. Chen, M. Mancini, X. Zhu и Z. Akata, „Semi-Supervised and Unsupervised Deep Visual Learning: A Survey,“ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, том 46, № 3, pp. 1327-1347, 2024. doi: <https://doi.org/10.1109/TPAMI.2022.3201576>.
- [53] Г. Агре, З. Марков и Д. Дочев, „Бейсов класификатор,“ в *Увод в машинното самообучение*, Софтех, 2001.
- [54] E. Gomedе, „Understanding Multinomial Naive Bayes Classifier,“ 11 November 2023. [Online]. Available: <https://medium.com/@evertongomede/understanding-multinomial-naive-bayes-classifier-fdbd41b405bf>. [Accessed 31 December 2023].
- [55] Г. Агре, З. Марков и Д. Дочев, „Машина на поддържащи вектори,“ в *Увод в машинното самообучение*, Софтех, 2001.
- [56] 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>.
- [57] Y. Wang, H. Lin, L. She, C. Li, L. Sun and J. Wang, „The Design and Implementation of Information-Based Teaching Skills Training Platform for Primary and Secondary School Teachers,“ in *International Conference on Education, Network and Information Technology (ICENIT)*, 2022. doi: 10.1109/ICENIT57306.2022.00017.
- [58] S. Mishra, „Breaking Down the Support Vector Machine (SVM) Algorithm,“ 29 October 2020. [Online]. Available: <https://towardsdatascience.com/breaking-down-the-support-vector-machine-svm-algorithm-d2c030d58d42>. [Accessed 30 December 2023].
- [59] 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>.
- [60] 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>.
- [61] 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.
- [62] 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, № 1, 2024. doi: <https://doi.org/10.1186/s40537-024-00973-y>.
- [63] B. Sun и H. Chen, „A Survey of k Nearest Neighbor Algorithms for Solving the Class Imbalanced Problem,“ *Wireless Communications and Mobile Computing*, 2021. doi: <https://doi.org/10.1155/2021/5520990>.

- [64] Г. Агре, З. Марков и Д. Дочев, „Основни методи за нейерархична и йерархична клъстеризация,“ в *Увод в машинното самообучение*, Софтех, 2001.
- [65] 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>.
- [66] J. Raymaekers and R. H. Zamar, “Regularized k-means through hard-thresholding,” *The Journal of Machine Learning Research*, vol. 23, no. 93, p. 1–48, 2022. doi: <https://doi.org/10.48550/arXiv.2010.00950>.
- [67] D. Wei and Z. Zhang, “K-means Clustering Algorithm based on Improved Density Peak,” in *BIC '23: Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing*, 2023. doi: <https://doi.org/10.1145/3592686.3592706>.
- [68] I. K. Jabari, S. Shofiyah, P. K. Sugiharto, N. N. Putriwijaya and N. Yudistira, “Learning-Augmented K-Means Clustering Using Dimensional Reduction,” in *SIET '23: Proceedings of the 8th International Conference on Sustainable Information Engineering and Technolo*, 2023. doi: <https://doi.org/10.1145/3626641.3627239>.
- [69] W. Liu, N. Wang and Y. Huang, “Outlier Detection Method based on Improved K-means Clustering Algorithm,” in *EITCE '21: Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 2021. doi: <https://doi.org/10.1145/3501409.3501648>.
- [70] K. Makarychev and L. Shan, “Explainable k-means: don't be greedy, plant bigger trees,” in *STOC 2022: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 2023. doi: <https://doi.org/10.1145/3519935.3520056>.
- [71] T. Yu и К. Т. Nwet, „Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System,“ в *ICCDE '20: Proceedings of 2020 6th International Conference on Computing and Data Engineering*, 2020. doi: <https://doi.org/10.1145/3379247.3379293>.
- [72] 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.
- [73] Y. Bao, H. Lu и Q. Gao, „Distance-based Fuzzy K-means Clustering without Cluster Centroids,“ 2024. doi: <https://doi.org/10.48550/arXiv.2404.04940>.
- [74] X. Geng, Y. Mu, S. Mao, J. Ye и L. Zhu, „An Improved K-Means Algorithm Based on Fuzzy Metrics,“ *IEEE Access*, том 8, pp. 217416-217424, 2020. doi: 10.1109/ACCESS.2020.3040745.
- [75] Y. Li и X. Xie, „Deep multi-view fuzzy k-means with weight allocation and entropy regularization,“ *Applied Intelligence*, том 53, № 24, pp. 30593-30606, 2023. doi: <https://doi.org/10.1007/s10489-023-05113-2>.
- [76] 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>.
- [77] D. Sanz-Alonso и R. Yang, „Unlabeled data help in graph-based semi-supervised learning: a Bayesian nonparametrics perspective,“ *The Journal of Machine Learning Research*, том 23, № 1, pp. 4205 - 4232, 2022.
- [78] Y. Ouali, C. Hudelot и M. Tami, „An Overview of Deep Semi-Supervised Learning,“ *arXiv:2006.05278*, 2020. doi: 10.48550/arXiv.2006.05278.
- [79] S. K. Mishra, A. Tripathi и M. Chauhan, „The role of semi-supervised learning in harnessing unlabeled data for model training,“ *Pharma innovation*, том 8, № 4, 2019. doi: <https://doi.org/10.22271/tpi.2019.v8.i4Sa.25255>.
- [80] X. Yang, Z. Song, I. King и Z. Xu, „A Survey on Deep Semi-Supervised Learning,“ *IEEE Transactions on Knowledge and Data Engineering*, том 35, № 10, pp. 8934-8954, 2023. doi: 10.1109/TKDE.2022.3220219.
- [81] J. Kim, S. Park, S.-D. Roh и K.-S. Chung, „An Efficient Noisy Label Learning Method with Semi-supervised Learning: An Efficient Noisy Label Learning Method with Semi-supervised Learning,“ в *ICMVA '23: Proceedings of the 2023*

- 6th International Conference on Machine Vision and Applications, 2023. doi: <https://doi.org/10.1145/3589572.3589596>.
- [82] A. Mey и M. Loog, „Improved Generalization in Semi-Supervised Learning: A Survey of Theoretical Results,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, том 45, № 4, pp. 4747 - 4767, 2022. doi: 10.1109/TPAMI.2022.3198175.
 - [83] P. K. Mvula, P. Branco, G.-V. Jourdan и H. L. Viktor, „A Survey on the Applications of Semi-supervised Learning to Cyber-security,” *ACM Computing Surveys*, том 56, № 10, pp. 1-41, 2024. doi: <https://doi.org/10.1145/3657647>.
 - [84] J. M. Duarte и L. Berton, „A review of semi-supervised learning for text classification,” *Artificial Intelligence Review*, том 56, № 9, pp. 9401 - 9469, 2023. doi: <https://doi.org/10.1007/s10462-023-10393-8>.
 - [85] Q. Zhu, L. Hu и Z. Xu, „Semi-supervised image classification based on deep clustering and pre-trained models,” в *CISAI '24: Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*, 2024. doi: <https://doi.org/10.1145/3703187.3703234>.
 - [86] J. Wu, J. Pang и Q. Huang, „Feature-based Perturbation Makes a Better Ensemble Learning for SSL Classification,” в *CVIPPR '24: Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recogniti*, 2024. doi: <https://doi.org/10.1145/3663976.3664035>.
 - [87] Z. Liang, Y.-G. Wang, W. Lu и X. Cao, „Boosting Semi-Supervised Learning with Dual-Threshold Screening and Similarity Learning,” *ACM Transactions on Multimedia Computing, Communications and Applications*, том 20, № 9, pp. 1-19, 2024. doi: <https://doi.org/10.1145/3672563>.
 - [88] Z. W. a. Z. S. Z. Zhang, „An improved algorithm of TFIDF combined with Naive Bayes,” в *7th International Conference on Multimedia and Image Processing (ICMIP '22)*, 2022. doi: <https://doi.org/10.1145/3517077.3517104>.
 - [89] A. Kumar, S. Ghosh и J. Verma, „Guided Self-Training based Semi-Supervised Learning for Fraud Detection,” в *ICAIF '22: Proceedings of the Third ACM International Conference on AI in Finance*, 2022. doi: <https://doi.org/10.1145/3533271.3561783>.
 - [90] Y. Zhuang, J. Song, N. Sadagopan и A. Beniwal, „Self-supervised Pre-training and Semi-supervised Learning for Extractive Dialog Summarization,” в *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*, 2023. doi: <https://doi.org/10.1145/3543873.3587680>.
 - [91] M. Fouad, W. Hussein, S. Rady, P. S Yu and Ghar, “Hybrid Recommender System Combining Collaborative Filtering with Utility Mining,” *International Journal of Intelligent Computing and Information Sciences*, vol. 22, no. 4, pp. 13-24, 2022. doi: 10.21608/ijicis.2022.145103.1192.
 - [92] P. Jain, “Basics of CountVectorizer,” 24 May 2021. [Online]. Available: <https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c>. [Accessed 28 November 2023].
 - [93] K. Ganesan, “HashingVectorizer vs. CountVectorizer,” [Online]. Available: <https://kavita-ganesan.com/hashtingvectorizer-vs-countvectorizer/>. [Accessed 29 November 2023].
 - [94] A. Rajaraman and J. D. Ullman, “Data Mining,” in *Mining of Massive Datasets*, Cambridge University Press, 2012. doi: <https://doi.org/10.1017/CBO9781139058452>.
 - [95] 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/>.
 - [96] K. Li and C. Fan, “Research and Application on Text Classification Model Based on Keywords,” in *CSAE '21: Proceedings of the 5th International Conference on Computer Science and Application Engineering*, 2021. doi: <https://doi.org/10.1145/3487075.3487159>.
 - [97] M. Chaudhary, “TF-IDF Vectorizer scikit-learn,” 24 April 2020. [Online]. Available: <https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a>. [Accessed 20 December 2023].

- [98] 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.
- [99] 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.
- [100] 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.
- [101] D. Suresh Asudani, N. K. Nagwani и P. Singh, „Impact of word embedding models on text analytics in deep learning environment: a review,“ *Artificial Intelligence Review*, том 56, № 9, p. 10345–10425, 2023. doi: <https://doi.org/10.1007/s10462-023-10419-1>.
- [102] 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.
- [103] C. Zhang, B. Peng, X. Sun, Q. Niu, J. Liu, K. Chen, M. Li, P. Feng, Z. Bi, M. Liu, Y. Zhang, X. Song, C. Fei, C. Heqi Yin, L. K. Yan, H. He и T. Wang, „From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models,“ *arXiv*, 2024. doi: 10.48550/arXiv.2411.05036.
- [104] J. Golec и T. Hachaj, „Ten Natural Language Processing Tasks with Generative Artificial Intelligence,“ *Applied Sciences*, том 15, № 16, 2025. doi: <https://doi.org/10.3390/app15169057>.
- [105] H. Alkaabi, A. K. Jasim и A. Darroudi, „From Static to Contextual: A Survey of Embedding Advances in NLP,“ *PERFECT: Journal of Smart Algorithms*, том 2, № 2, pp. 64-73, 2025. doi: 10.62671/perfect.v2i2.77.
- [106] D. Jurafsky и J. H. Martin, *Speech and language processing*, Stanford University, 2024.
- [107] 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.
- [108] J. Neidhardt, „Transforming recommender systems: Balancing personalization, fairness, and human values.,“ в *IJCAI '24: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. doi: <https://doi.org/10.24963/ijcai.2024/982>.
- [109] M. Zhang, „A Novel Implementation of Library Personalized Recommendation Service based on Associative Rule Mining Model,“ в *International Conference on Intelligent Computing and Knowledge Extraction (ICICKE)*, 2025. doi: 10.1109/ICICKE65317.2025.11136659.
- [110] A. A. Adewojo и A. O. Dunmade, „From big data to intelligent libraries: Leveraging analytics for enhanced user experiences,“ *Business Information Review*, том 41, № 3, pp. 104-109, 2024.
- [111] X. Wu и G. Sang, „An Accurate Recommendation Method of University Library Digital Resources Based on User Interest,“ *International Journal of High Speed Electronics and Systems*, том 35, № 3, 2025. doi: 10.1142/S0129156425404528.
- [112] N. Gulati, S. Sharma и S. Tyagi, „Transforming digital libraries through AI: analysing personalization strategies for improved user satisfaction,“ *ShodhKosh: Journal of Visual and Performing Arts*, том 5, № 5, 2024. doi: 10.29121/shodhkosh.v5.i5.2024.2077.
- [113] U. F. Ikwuanusi, P. A. Adepoju и C. S. Odionu, „AI-driven solutions for personalized knowledge dissemination and inclusive library user experiences,“ *International Journal of Engineering Research Updates*, том 4, № 2, pp. 052-062, 2023. doi: 10.53430/ijeru.2023.4.2.0023.
- [114] S. S. Suryawanshi, D. B. Deshmukh и M. Jamdade, „Reimagining library services: The role of artificial intelligence in modern library management,“ *International Scientific Journal of Engineering and Management*, том 4, № 8, pp. 1-9, 2025. doi: 10.55041/ISJEM.PKDAL002.

- [115] M. M. Rahman, F. F. Sifat, R. M. Islam, S. Molla и M. R. K. Khan, „Hybrid recommendation systems using adaptive clustering to address cold start problems,” *International Conference on Electrical, Computer and Energy Technologies*, pp. 1-6, 2024. doi: 10.1109/ICECET61485.2024.
- [116] S. S. Patil, L. Y. Kamble и P. Bagewadi, „Transformative role of artificial intelligence in library management: A review,” *The Journal of Library and Information Management*, том 16, № 1, pp. 27-40, 2025.
- [117] O. A. S. Ibrahim, E. M. G. Younis, E. A. Mohamed и W. N. Ismail, „Revisiting recommender systems: an investigative survey,” *Neural Comput & Applic*, том 37, p. 2145–2173, 2025. doi: <https://doi.org/10.1007/s00521-024-10828-5>.
- [118] C. Troussas, A. Krouska, A. Koliarakis и C. Sgouropoulou, „Harnessing the power of user-centric artificial intelligence: Customized recommendations and personalization in hybrid recommender systems,” *Artificial Intelligence Models, Tools and Applications with A Social and Semantic Impact*, том 12, № 5, 2023. doi: <https://doi.org/10.3390/computers12050109>.
- [119] E. Purificato, L. Boratto и E. W. D. Luca, „User modeling and user profiling: A comprehensive survey,” *arXiv:2402.09660*, 2024. doi: <https://doi.org/10.48550/arXiv.2402.09660>.
- [120] Y. Liu, Y. Xu и S. & Zhou, „Enhancing user experience through machine learning-based personalized recommendation systems: Behavior data-driven UI design,” *Applied and Computational Engineering*, том 112, № 1, pp. 42-46, 2024.
- [121] F. A. Najla, D. S. Kusumo и A. Gandhi, „Interaction Design for Book Recommendation System to Improve User Satisfaction at EPerpusdikbud,” в *IEEE International Conference on Data and Software Engineering (ICoDSE)*, 2023. doi: 10.1109/ICoDSE59534.2023.10292072.
- [122] Z. Liao, L. Chen, Y. Qi и F. Li, „DPBD: Disentangling Preferences via Borrowing Duration for Book Recommendation,” *Big Data and Cognitive Computing*, том 9, № 9, 2025. doi: <https://doi.org/10.3390/bdcc9090222>.
- [123] M. S. Hidri, „Learning-based models for building user profiles for personalized information access,” *Interdisciplinary Journal of Information, Knowledge, and Management*, том 19, 2024. doi: <https://doi.org/10.28945/5275>.
- [124] Z. D. Champiri, B. Fisher, L. C. Kiong и M. Danaee, „How Contextual Data Influences User Experience with Scholarly Recommender Systems: An Empirical Framework,” в *HCI International 2020 - Late Breaking Papers: User Experience Design and Case Studies: 22nd HCI International Conference, HCII 2020*, Copenhagen, Denmark, 2020. doi: https://doi.org/10.1007/978-3-030-60114-0_42.
- [125] S. Unnisa и N. S. Akshaya, „Personalized mood-centric book recommendation integrating machine learning with content based filtering,” *International Journal of Information Technology, Research and Applications*, том 3, № 3, pp. 15-22, 2024. doi: 10.59461/ijitra.v3i3.100.
- [126] A. Singh, S. P. Gandhala, M. Alahmar, P. Gaikwad, U. Wable, A. Yadav и R. Agrawal, „Towards Context Aware and Age-Based Book Recommendation System using Machine Learning for Young Readers,” в *Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2024. doi: 10.1109/ICAECT60202.2024.10469429.
- [127] Y. Liang и J. Wang, „Intelligent library recommendation based on GNN and attention networks,” *Journal of Computational Methods in Sciences and Engineering*, 2025. doi: <https://doi.org/10.1177/14727978251364472>.
- [128] T. T. Lin и Y. Li, „Recommender System Powered by Large Language Models,” в *16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2024. doi: 10.1109/IIAI-AAI63651.2024.00092.
- [129] J. Liu, „Application of Machine Learning Book Recommendation Strategy Based on User Reading Preferences in Smart Libraries,” в *International Conference on Telecommunications and Power Electronics (TELEPE)*, 2024. doi: 10.1109/TELEPE64216.2024.00007.
- [130] J. Liu, „Design of Book Recommendation System Based on Machine Learning in Smart Library,” в *3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, 2024. doi: 10.1109/AIARS63200.2024.00016.

- [131] T. Pan, „Personalized Recommendation Service in University Libraries using Hybrid Collaborative Filtering Recommendation System,” в *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, 2024. doi: 10.1109/IACIS61494.2024.10721676.
- [132] Y. Kong, „Enhancing library digitalization: A heterogeneous network embedding approach for personalized book recommendations,” *IEEE Access*, том 12, pp. 188723-188738, 2024. doi: 10.1109/ACCESS.2024.3418957.
- [133] I. Sanjaya, A. Sujada и Y. H. C. Pratama, „Implementation of content-based filtering in a novel recommendation system to enhance user experience,” *bit-Tech*, том 8, № 1, 2025. doi: 10.32877/bt.v8i1.2833.
- [134] S. S. Kumar и S. Barathkumar, Enhancing book discovery with collaborative filtering and machine learning, 2024. doi: <https://doi.org/10.1201/9781003543633-33>.
- [135] P. Y. Ravish и V. G. Kale, „Transforming libraries with artificial intelligence: Applications, benefits, and future trends,” *International Scientific Journal of Engineering and Management*, том 4, № 8, pp. 1-9, 2025. doi: 10.55041/ISJEM04979.
- [136] Q. Wang и Q. Chen, „Personalized book recommendation based on relational graph convolutional network,” в *10th International Conference on Big Data and Information Analytics (BigDIA)*, 2024. doi: 10.1109/BigDIA63733.2024.10808689.
- [137] S. Xie и Z. Liu, „Library resource recommendation integrating collaborative filtering and association rules,” *Journal of Computational Methods in Sciences and Engineering*, 2025. doi: <https://doi.org/10.1177/14727978251364442>.
- [138] R. Sivasankari, S. Suriya, S. Sindhu, J. Devi и J. Dhilipan, „AI-powered recommendation systems and resource discovery for library management,” в *Applications of Artificial Intelligence in Libraries*, 2024, pp. 223-244. doi: <https://doi.org/10.4018/979-8-3693-1573-6.ch009>.
- [139] E. Mupaikwa, „The application of artificial intelligence for reference purposes in academic libraries,” в *Applications of Artificial Intelligence in Libraries*, 2024, pp. 166-192. doi: <https://doi.org/10.4018/979-8-3693-1573-6.ch007>.
- [140] X. Wang и L. Lui, „The role and function of artificial intelligence and the metaverse in smart libraries,” *Applied mathematics and nonlinear sciences*, том 9, № 1, 2024. doi: 10.2478/amns-2024-1578.
- [141] F. D. Lorenzis, A. Visconti, A. Cannavo и F. Lamberti, „MetaLibrary: Towards Social Immersive Environments for Readers,” в *Extended Reality: International Conference*, 2023. doi: https://doi.org/10.1007/978-3-031-43404-4_6.
- [142] H. Sa'ari, M. D. Sahak и S. Skrzyszewskis, „Deep Learning Algorithms for Personalized Services and Enhanced User Experience in Libraries,” *Mathematical Sciences and Informatics Journal*, том 4, № 1, pp. 30-47, 2023. doi: 10.24191/mij.v4i2.23026.
- [143] M. D. Mpela и T. Zuva, „Deep Learning-Based Long and Short-Term Memory Integration for Enhanced Library Recommendation Systems,” в *In 2024 4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2024. doi: 10.1109/IMITEC60221.2024.10851043.
- [144] N. Pang, „A Personalized Recommendation Algorithm for Semantic Classification of New Book Recommendation Services for University Libraries,” *Mathematical Problems in Engineering*, pp. 1-8, 2022. doi: <https://doi.org/10.1155/2022/8740207>.
- [145] Y. Cai, „Core technologies in recommender systems: Investigating and analyzing standard implementations,” *Highlights in Science Engineering and Technology*, том 94, pp. 259-272, 2024. doi: 10.54097/thv9yp09.
- [146] X. Li, „Research on the Application of Artificial Intelligence in Library Reader Behavior Analysis and Personalized Service,” в *2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, 2024. doi: 10.1109/CIPAE64326.2024.00171.
- [147] V. Pawar, „Book recommendation system with integrated chatbot,” *Indian Scientific Journal Of Research In Engineering And Management*, том 8, № 5, pp. 1-5, 2024.

- [148] H. Mojada и S. Sivakrishna, „Artificial intelligence based academic libraries: A review,“ в *Futuristic Trends in Artificial Intelligence*, том 3, 2024. doi: 10.58532/V3BIAI12P5CH.
- [149] R. Visnudharshana и H. S. Kishore, „AI-Driven Language Enhancement Strategies for Libraries: Empowering Information Access and User Experience in an English Language Context,“ в *Improving Library Systems with AI: Applications, Approaches, and Bibliometric Insights*, 2024. doi: 10.4018/979-8-3693-5593-0.ch018, pp. 244-253.
- [150] A. Kavak, „Integration of artificial intelligence (AI) technologies into academic library services: user opinions in Türkiye,“ *Reference Services Review*, 2024. doi: 10.1108/RSR-09-2024-0054.
- [151] I. Tai и S. Ghosh, „Integrating AI into Library Systems: A Perspective on Applications and Challenges,“ *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries. Association for Computing Machinery*, том 1, № 11, 2025. doi: <https://doi.org/10.1145/3677389.3702568>.
- [152] „Personalized book recommendation algorithm for university library based on deep learning models,“ *Journal of Sensors*, pp. 1-6, 2022.
- [153] F. D. Hengst, E. M. Grua, A. E. Hassouni и M. Hoogendoorn, „Reinforcement learning for personalization: A systematic literature review,“ *Data Science*, том 3, № 10, pp. 1-41, 2020. doi: 10.3233/DS-200028.
- [154] A. Alomran и I. Basha, „An AI-based classification and recommendation system for digital libraries,“ *Scalable Computing: Practice and Experience*, том 25, № 4, pp. 3181-3199, 2024. doi: <https://doi.org/10.12694/scpe.v25i4.2882>.
- [155] V. S. Kumaran и R. H. Latha, „Towards personal learning environment by enhancing adaptive access to digital library using ontology-supported collaborative filtering,“ *Library Hi Tech*, том 41, № 6, p. 1658–1675, 2023. doi: 10.1108/LHT-12-2021-0433.
- [156] A. Koliarakis, A. Krouska, C. Troussas и C. Sgouropoulou, „Modified collaborative filtering for hybrid recommender systems and personalized search: The case of digital library,“ в *17th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*, 2022. doi: 10.1109/SMAP56125.2022.9942020..
- [157] M. D. Mpela и T. Zuva, „Deep Learning-Based Long and Short-Term Memory Integration for Enhanced Library Recommendation Systems,“ в *4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2024.
- [158] T. Xue, X. Wang и T. Liu, „Research on Personalized Recommendation System of Library Collection Based on Deep Learning,“ в *2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, 2024. doi: 10.1109/ICIPCA61593.2024.10709150.
- [159] F. Wang, „Personalized Recommendation Algorithm with Data Mining for College Libraries using Long Short Term Memory Technique,“ в *Second International Conference on Data Science and Information System (ICDSIS)*, 2024. doi: 10.1109/ICDSIS61070.2024.
- [160] M. G. Nitu, M. Dascalu, M. D. Dascalu, L. Neagu и M. Dascalu, „Lib2life – digital library services empowered with advanced natural language processing techniques,“ *Interaction Design and Architecture(s)*, 2024. doi: 10.55612/s-5002-060-006.
- [161] H. Xue, K. Guo и F. He, „Research on Personalized Recommendation Service of Mobile Library Based on User Portrait,“ в *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2022. doi: 10.1109/ICAICA54878.2022.9844517.
- [162] Z. D. Champiri, B. Fisher, L. C. Kiong и M. Danaee, „How Contextual Data Influences User Experience with Scholarly Recommender Systems: An Empirical Framework,“ *HCI International 2020 - Late Breaking Papers: User Experience Design and Case Studies. HCII 2020*, том 12423, 2021. doi: https://doi.org/10.1007/978-3-030-60114-0_42.
- [163] S. Shee, „Using Artificial Intelligence in Academic Library Services: Transforming the Future of Learning,“ *Science, Architecture, Technology and Environment*, том 2, № 4, 2025. doi: 10.63680/nklcf5435q.

- [164] Y. Liu, „Survey of intelligent recommendation of academic information in university libraries based on situational perception method,” *Journal of Education and Learning*, том 9, № 2, pp. 197-202, 2020. doi: 10.5539/jel.v9n2p197.
- [165] M. Zanker, L. Rook и D. Jannach, „Measuring the impact of online personalisation: Past, present and future,” *International Journal of Human-Computer Studies*, том 131, № 3, pp. 160-168, 2019. doi: 10.1016/j.ijhcs.2019.06.006.
- [166] „Дигитална библиотека Народна библиотека "Иван Вазов" Пловдив,” [Онлайн]. Available: <https://digital.libplovdiv.com/bg>.
- [167] Supriyono, A. P. Wibawa, Suyono и F. Kurniawan, „Advancements in natural language processing: Implications, challenges, and future directions,” *Telematics and Informatics Reports*, том 16, № 5, 2024. doi: 10.1016/j.teler.2024.100173.
- [168] T. Ha and X. Gao, “Evolving Multi-view Autoencoders for Text Classification,” in *WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, New York, 2022. doi: <https://doi.org/10.1145/3486622.3493969>.
- [169] 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>.
- [170] M. Ashraf, F. Anowar, J. H. Setu, A. I. Chowdhury, E. Ahmed и A. Islam, „A Survey on Dimensionality Reduction Techniques for Time-Series Data,” *IEEE Access*, том 11, pp. 42909-42923, 2023. doi: 10.1109/ACCESS.2023.3269693.
- [171] Y.-c. I. Chang , „A Survey: Potential Dimensionality Reduction Methods,” *arXiv:2502.11036*, 2025. doi: <https://doi.org/10.48550/arXiv.2502.11036>.
- [172] L. McInnes, J. Healy и J. Melville, „UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.
- [173] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger и Y. Kluger, „Fast interpolation-based t-SNE for improved visualization of large datasets,” *Nature Methods*, том 16, № 3, p. 243–245, 2019. doi: 10.1038/s41592-018-0308-4.
- [174] F. Heimerl, S. Koch, T. Kronqvist, G. Shadow и O. Deussen, „Visual comparison of dimensionality reduction techniques for text data,” *Computer Graphics Forum*, том 39, № 3, p. 87–100, 2020.
- [175] Y. Wang, H. Yao и S. Zhao, „Auto-encoder based dimensionality reduction,” *Neurocomputing*, 2016. doi: 10.1016/j.neucom.2015.08.104.
- [176] Z. Yang, Z. Hu, R. Salakhutdinov и T. Berg-Kirkpatrick, „Improved variational autoencoders for text modeling using dilated convolutions,” в *In Proceedings of the 34th International Conference on Machine Learning*, 2017.
- [177] 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>.
- [178] 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.
- [179] 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>.
- [180] 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>.
- [181] 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>.

- [182] 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>.
- [183] R. K. Merton, „The Matthew Effect in Science," *Science*, том 159, № 3810, p. 56–63, 1968.
- [184] E. U. A. f. N. a. I. S. (ENISA), „Guidelines for SMEs on the Security of Personal Data Processing," 2016.
- [185] E. U. A. f. N. a. I. S. (ENISA), „Handbook on Security of Personal Data Processing," ENISA, 2017.
- [186] 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>.
- [187] „Sentence Transformers," [Онлайн]. Available: <https://huggingface.co/sentence-transformers>.
- [188] „Machine Learning in Python," [Online]. Available: <https://scikit-learn.org/stable/>.
- [189] „scikit-fuzzy," [Онлайн]. Available: <https://pypi.org/project/scikit-fuzzy/>.
- [190] A. Barbaresi, „Simplemma: a simple multilingual lemmatizer for Python," [Онлайн]. Available: <https://github.com/adbar/simplemma>.
- [191] „Stopwords Bulgarian (BG)," [Онлайн]. Available: <https://github.com/stopwords-iso/stopwords-bg>.
- [192] „NumPy," [Онлайн]. Available: <https://pypi.org/project/numpy/>.
- [193] „SciPy," [Онлайн]. Available: <https://scipy.org/>.
- [194] „PyTorch," [Онлайн]. Available: <https://pytorch.org/>.
- [195] „Hugging Face," [Онлайн]. Available: <https://huggingface.co/>.
- [196] Davlan, „distilbert-base-multilingual-cased-ner-hrl [Large language model]," 2023. [Онлайн]. Available: <https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl>.
- [197] rmihaylov, „bert-base-ner-theseus-bg [Large language model]," 2024. [Онлайн]. Available: <https://huggingface.co/rmihaylov/bert-base-ner-theseus-bg>.
- [198] Babelscape, „wikineural-multilingual-ner [Large language model]," 2021. [Онлайн]. Available: <https://huggingface.co/Babelscape/wikineural-multilingual-ner>.
- [199] auhide, „bert-bg-ner [Large language model]," [Онлайн]. Available: <https://huggingface.co/auhide/bert-bg-ner>.
- [200] B. Group. [Онлайн]. Available: <https://github.com/omwn/omw-data/tree/main/wns/bul>.
- [201] T. A. Davis, Direct methods for sparse linear systems, Society for Industrial and Applied Mathematics, 2006.
- [202] K. Järvelin и J. Kekäläinen, „umulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, том 20, № 4, p. 422–446, 2002. doi: <https://doi.org/10.1145/582415.582418>.

ДЕКЛАРАЦИЯ ЗА ОРИГИНАЛНОСТ НА РЕЗУЛТАТИТЕ

Декларирам, че настоящият дисертационен труд на тема: „Методи и алгоритми за персонализация и адаптивност в среди за управление на съдържание“, с научен ръководител: проф. д-р Десислава Панева-Маринова, съдържа оригинални резултати, получени при извършена от мен изследователска дейност. Всички използвани резултати от чужди автори и източници са надлежно и точно цитирани в библиографията.

Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Емануела Митрева

