**Документация, содержащая описание функциональных характеристик модуля программного обеспечения**

**«Поисковый модуль для стриминговых сервисов»**

**г. Москва 2025**

**Содержание**

[1. Описание модуля «Поисковый модуль для стриминговых сервисов» 3](#_s2xpnm9x5zoo)

[2. Описание функциональных характеристик 3](#_7ep7zxhbxr7i)

[3. Архитектура модуля «Персонализация контента» 4](#_fuxsqccpbuer)

[4. Схема работы 6](#_spk4gfqhsfnh)

[5. Используемые языки программирования и фреймворки 10](#_pf377y6aljju)

[6. Описание технической инфраструктуры](#_9nne77bsuqky) 11

# **Описание модуля «Поисковый модуль для стриминговых сервисов»**

Поисковый модуль для стриминговых сервисов — это технологический инструмент, который использует алгоритмы и методы анализа данных, машинного обучения и обработки естественного языка (NLP) для быстрого и точного поиска контента, соответствующего запросам пользователей. Основная цель системы — обеспечить удобный и эффективный доступ к разнообразному контенту, включая пользовательский контент (UGC, user-generated content), фильмы, сериалы, музыкальные видео и другие материалы.

Поисковый модуль ориентирован на обработку запросов в условиях большого объема данных и разнообразия контента. Он учитывает как текстовые запросы, так и контекстные данные, такие как популярность контента, релевантность, предпочтения пользователя и текущие тренды. Это позволяет системе предлагать наиболее подходящие результаты, даже если запросы формулируются неточно или содержат опечатки.

# **Описание функциональных характеристик**

2.1 Поиск по разным типам сущностей: Поддержка поиска по различным сущностям, таким как каналы, видео, персонажи и другие элементы контента. Это позволяет пользователям находить не только видео, но и другие типы контента, например, авторов или каналы.

2.2 Поиск по разным типам контента: Модуль поддерживает поиск по различным типам контента, таким как пользовательский контент (UGC), фильмы, сериалы, музыкальные клипы, короткие видео и другие виды материалов. Это обеспечивает гибкость и точность в поиске различных видов контента.

2.3 Поиск по новому контенту: Включает функциональность быстрого поиска нового контента, например, трансляций или видео, которые только что начали или только что были загружены. Это позволяет пользователям всегда быть в курсе актуальных и свежих материалов.

2.4 Поддержка сложных запросов: Модуль способен обрабатывать длинные и сложные запросы, позволяя фильтровать и сортировать результаты по категориям, жанрам, дате публикации и другим параметрам. Это предоставляет пользователям больше возможностей для уточнения своих запросов.

2.5 Ранжирование результатов: Поисковая система учитывает популярность, релевантность и качество контента при ранжировании результатов поиска. Это гарантирует, что наиболее подходящие материалы будут отображаться в верхней части списка результатов, а менее релевантные — внизу.

2.6 Автодополнение и исправление ошибок: Система поддерживает автодополнение запросов по мере ввода текста и автоматически исправляет опечатки, чтобы обеспечить точность поиска, даже если запрос введен с ошибками.

2.7 Персонализация: Модуль учитывает историю поиска и предпочтений пользователя для повышения точности и релевантности результатов. Это позволяет системе предлагать контент, который соответствует интересам и предпочтениям конкретного пользователя.

2.8 Транслитерация: Модуль учитывает транслитерацию и возможные ошибки, связанные со сменой раскладки клавиатуры, чтобы улучшить качество поиска и сделать его более точным для пользователей, вводящих запросы на разных языках или в разных раскладках.

# **Архитектура** модуля «Поисковый модуль для стриминговых сервисов»

В рамках интеграции поискового модуля для стриминговых сервисов с крупной медиа платформой активно используются различные внешние сервисы, которые обеспечивают функциональность, необходимую для эффективной работы системы поиска и взаимодействия с пользователями. Включение этих сервисов в архитектуру поискового модуля позволяет не только повысить качество поиска, но и улучшить пользовательский опыт за счет персонализации, аналитики и оптимизации.

***Внешние сервисы, участвующие в интеграции***

Авторизация: Сервис для регистрации, авторизации и аутентификации пользователей. Этот сервис обеспечивает безопасность и доступность персонализированного контента для каждого пользователя. Он позволяет уникально идентифицировать пользователя, а также управлять его доступом к различным материалам на платформе.

A/B тестирование: Сервис, который разделяет пользователей на группы для проведения A/B тестов. Это позволяет экспериментировать с различными вариантами интерфейса или функционала поискового модуля, чтобы выбрать наиболее эффективный для улучшения пользовательского опыта и повышения вовлеченности.

Бэкенд сервисы: Ряд различных бэкенд сервисов, которые обеспечивают работу всей видео платформы. Эти сервисы взаимодействуют с поисковым модулем для обеспечения правильного отображения результатов поиска, предоставляя пользователю доступ к видео, каналам и другому контенту. Также эти сервисы играют ключевую роль в обработке запросов и обеспечении масштабируемости системы.

Логи плеера: Сервис для сбора логов (событий) видео плеера в режиме реального времени. Этот сервис необходим для отслеживания пользовательских действий, таких как паузы, воспроизведение, переходы по видео и другие важные события. Эти данные позволяют улучшить точность персонализированных рекомендаций и анализировать поведение пользователей.

AppMetrica: Сервис для сбора пользовательских событий в мобильных приложениях. AppMetrica помогает собирать подробные данные о взаимодействии пользователей с мобильным приложением, включая клики, просмотры контента, а также другие действия, которые могут быть использованы для улучшения поиска и персонализации результатов.

Яндекс.Метрика: Сервис для сбора пользовательских событий в web приложении (Desktop и мобильный Web). Яндекс.Метрика предоставляет подробную информацию о поведении пользователей в web-версии платформы, включая анализ запросов, популярности контента и других ключевых показателей. Эти данные используются для улучшения поисковых алгоритмов и предложения более релевантного контента.

***Интеграция внешних сервисов с поисковым модулем***

Интеграция этих внешних сервисов с поисковым модулем помогает создать динамичную и гибкую систему, способную эффективно работать с большим объемом данных и запросов пользователей.

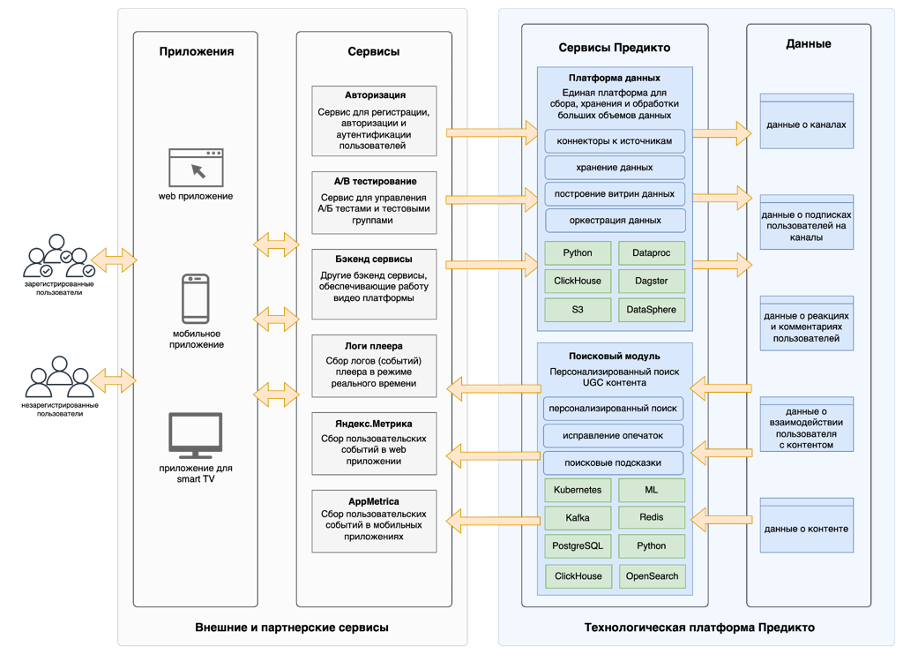
Авторизация служит основой для персонализированного подхода, где каждый пользователь получает результаты поиска, адаптированные под его предпочтения и историю взаимодействия с контентом.

A/B тестирование позволяет тестировать различные версии поисковых алгоритмов и интерфейса, что способствует оптимизации пользовательского опыта и повышению эффективности работы поисковой системы.

Бэкенд сервисы предоставляют необходимые ресурсы для обработки запросов пользователей, поиска контента и взаимодействия с базами данных, а также обеспечивают поддержку высоких нагрузок на систему.

Логи плеера и AppMetrica обеспечивают сбор данных о поведении пользователей в реальном времени, что позволяет анализировать эффективность поиска и адаптировать результаты под изменения в пользовательских предпочтениях.

Яндекс.Метрика и AppMetrica используются для сбора статистики, анализа пользовательского опыта и выявления возможных узких мест в функциональности поисковой системы



*рис.1 - концептуальная архитектура Поискового модуля для стриминговых сервисов*

# **Схема работы**

4.1 Уровень приложений

Пользователи получают доступ к поисковому модулю через различные интерфейсные приложения: мобильное приложение, веб-приложение, а также через смарт-телевизоры и другие устройства. Эти интерфейсные приложения не являются частью поискового модуля, однако играют ключевую роль в обеспечении взаимодействия пользователя с системой поиска.

Взаимодействие между интерфейсными приложениями и поисковым модулем осуществляется через API. Это позволяет приложениям отправлять запросы к поисковому модулю, получать результаты поиска и отображать их пользователю. API служит как стандартный интерфейс для обмена данными между интерфейсными приложениями и ядром поисковой системы, обеспечивая гибкость и масштабируемость решения.

4.2 Уровень внешних сервисов

Для корректной работы поискового модуля, он интегрирован с рядом внешних сервисов, которые предоставляют дополнительные данные и функциональность. Эти сервисы взаимодействуют с платформой Предикто через стандартизированные протоколы и API.

4.3. Платформа данных

Для интеграции и работы с внешними сервисами используется Платформа данных, которая является ключевым компонентом инфраструктуры. Платформа данных отвечает за сбор, хранение и обработку больших объемов данных, поступающих из различных источников. Интеграция с внешними сервисами может осуществляться как по пакетной, так и по потоковой обработке данных, в зависимости от специфики источника.

Модуль Инфраструктура для сервиса (или Платформа данных) реализует все необходимые функции по получению данных, их хранению и дальнейшей обработке. Этот модуль также отвечает за обеспечение высокой доступности данных и их обработку в реальном времени.

4.4. Поисковый модуль для стриминговых сервисов

Основной функционал поискового модуля заключается в персонализированном поиске контента на базе собранных данных о пользователях и контенте, с использованием технологий машинного обучения для улучшения качества поиска.

Основные компоненты поискового модуля включают:

Kubernetes: Платформа для оркестрации контейнеров, которая управляет развертыванием и масштабированием компонентов поискового модуля.

ML: Модели машинного обучения для ранжирования результатов поиска и персонализации контента, что позволяет точно подбирать материалы в соответствии с предпочтениями пользователя.

OpenSearch: Мощный поисковый движок, используемый для поиска и сортировки релевантного контента для каждого запроса.

Redis: Используется для кэширования данных и запросов, ускоряя процессы поиска и улучшая отклик системы.

PostgreSQL: Основная база данных, которая используется для хранения данных о пользователях, контенте и их взаимосвязях.

ClickHouse: Система для хранения и анализа больших объемов логов поискового модуля, что позволяет анализировать поведение пользователей и улучшать результаты поиска.

Kafka: Шина данных для обработки потоковых данных, которая помогает передавать данные между различными сервисами платформы в реальном времени.

4.5. Уровень данных

В поисковом модуле используются несколько категорий данных, которые являются ключевыми для работы с контентом и пользовательскими предпочтениями:

Данные о контенте: Информация о видео, включая название, описание, продолжительность, дату публикации и другие метаданные.

Данные об авторах/каналах: Сведения о каналах, такие как название, описание и время создания, что помогает пользователям находить авторов или каналы с интересующим контентом.

Данные о реакциях пользователей: Лайки, комментарии и другие реакции пользователей на контент, что позволяет учитывать пользовательскую активность в процессе поиска.

Данные о подписках пользователей: Информация о том, на какие каналы или авторов подписаны пользователи, что позволяет учитывать предпочтения при формировании поисковых результатов.

Данные о взаимодействии пользователей с контентом: Сюда включаются просмотры, досмотры, перемотки и другие действия, которые помогают системе точнее предсказывать, что может заинтересовать пользователя.

# **Используемые языки программирования и фреймворки**

Язык программирования

Python: Основной язык разработки, который используется для реализации всех компонентов системы, включая обработку данных, машинное обучение и работу с API.

Библиотеки и фреймворки

5.1. Обработка данных

Polars: Быстрая и эффективная библиотека для обработки данных в формате таблиц, работающая с большими объемами данных.

Pandas: Один из самых популярных инструментов для работы с данными в табличной форме, обеспечивающий удобный интерфейс для анализа и манипуляции данными.

deltalake: Библиотека для работы с форматом Delta Lake, который используется для управления данными в виде больших файлов и эффективного выполнения запросов.

PyArrow: Библиотека для работы с Apache Arrow, которая используется для обеспечения быстрого и эффективного обмена данными между системами.

Psycopg: Пакет для взаимодействия с PostgreSQL, используемый для работы с реляционными базами данных.

SciPy: Библиотека для научных и инженерных вычислений, обеспечивающая широкий набор математических и статистических функций.

Numba: Библиотека для ускорения вычислений с помощью JIT-компиляции Python кода на языке машинных команд.

5.2. Модели машинного обучения

scikit-learn: Популярная библиотека для машинного обучения, которая предоставляет простой и эффективный инструментарий для построения моделей и анализа данных.

PyTorch: Гибкий фреймворк для создания и обучения нейронных сетей, который активно используется для глубокого обучения и искусственного интеллекта.

CatBoost: Библиотека для градиентного бустинга, которая оптимизирована для работы с категориальными данными.

XGBoost: Мощная библиотека для реализации алгоритмов градиентного бустинга, широко используемая для решения задач классификации и регрессии.

LightFBM: Фреймворк для быстрого градиентного бустинга, специально предназначенный для работы с большим объемом данных.

Implicit: Библиотека для создания рекомендательных систем с использованием матричной факторизации.

Optuna: Оптимизационная библиотека для автоматического подбора гиперпараметров моделей машинного обучения.

5.3. Обработка текста и NLP

PyMorphy2: Библиотека для морфологического анализа русского языка.

Levenshtein: Библиотека для вычисления расстояния Левенштейна, используется в задачах обработки текста и анализа строк.

Sentence Transformers: Библиотека для преобразования предложений в эмбеддинги, используется для задач обработки естественного языка (NLP).

5.4. Операции с графами

Dagster: Фреймворк для разработки и оркестрации рабочих процессов и данных, используемый для построения надежных и масштабируемых систем обработки данных.

5.5. Дополнительные инструменты

rbloom: Библиотека для реализации фильтров Блума, применяемая для эффективной проверки принадлежности элемента множеству, что используется в рекомендательных системах и для работы с большими данными.

# Описание технической инфраструктуры

Облачные сервисы и виртуализация

Yandex Cloud: Используется для аренды виртуальных серверов и хранения данных в Yandex Object Storage.

Виртуализация: Используются Yandex Compute Cloud и Docker для контейнеризации и оркестрации с помощью Kubernetes.

Базы данных и кэширование

PostgreSQL: Основная база данных для хранения информации о пользователях и контенте.

Redis: Кэширование данных и запросов.

OpenSearch: Поисковый движок для обработки запросов.

ClickHouse: Хранение и анализ логов.

Машинное обучение

Yandex Data Sphere и Apache Spark для обработки данных и обучения моделей.

Dagster для управления рабочими процессами и данными.

PyTorch, scikit-learn, XGBoost, CatBoost для машинного обучения и построения моделей.

Системы мониторинга.

Prometheus и Grafana для сбора и анализа метрик.

Grafana Loki для сбора и анализа логов.

Инструменты для обработки текста.

PyMorphy2: Морфологический анализ русского языка.

Levenshtein: Вычисление расстояния Левенштейна для улучшения обработки текстовых запросов.

Sentence Transformers: Преобразование предложений в эмбеддинги для улучшения поиска.

Безопасность. API интеграция с внешними сервисами через стандартизированные протоколы. Используется x-authorization-token для авторизации.

# 

# 