## Краткое описание системы

Программное обеспечение Content Security предназначено для интеллектуального анализа данных о различном цифровом контенте, например, о фильмах или сериалах, по названию при помощи поисковых систем общего назначения.

Сбор информации из открытых источников происходит автоматически, необходимо только предварительно настроить нужные параметры в системе.

## Функциональные возможности

Система Content Security предназначена для простого автоматического поиска ссылок в сети Интернет, на которых содержится нужный пользователю цифровой контент (например, фильмы или сериалы).

Программное обеспечение обладает следующими функциональными возможностями:

* автоматический поиск и сбор ссылок (URL), на которых размещен искомый цифровой контент;
* автоматический поиск по выбранному сайту;
* возможность заведения в систему цифрового контента по названию;
* автоматический анализ собираемых ссылок и их содержимого по требуемым пользователю признакам. ПО поддерживает широкий набор признаков, по которым ПО может анализировать ссылки;
* наличие API для удобной и простой интеграции с любыми подсистемами;
* наличие удобного веб-интерфейса для настройки ПО (добавление в поиск цифрового контента по названию, настройка признаков, по которым анализируются найденные ссылки, настройка механизма отправки найденных ссылок в вашу подсистему).

Ниже представлено более подробное описание функциональных возможностей сервиса.

2.1 Загрузка данных в сервис из внешних источников

В систему загружается массив данных, содержащий информацию о контенте:

* Наименование;
* Год выпуска;
* Режиссер;
* Актер;

Данные сохраняются в БД. На основе поступивших данных формируется список задач к выполнению. На основе данных задач в дальнейшем выполняется формирование запросов в поисковик и парсинг данных.

2.2 Процесс парсинга сайтов

Парсинг — это автоматизированный сбор и структурирование информации с сайтов. Система Content Security поддерживает 2 вида парсинга:

* Общий (включает в себя поиск по всем сайтам);
* Точечный (Точечный (или точный) парсинг осуществляется по определенным сайтам).

2.3 Формирование задач и запросов для парсинга

Задачи являются исходным элементом для парсинга. Соответственно, для каждого запроса в поисковик имеется своя задача. Задача на парсинг имеет привязку ко времени. То есть для одного и того же контента существует множество задач с определенной периодичностью. К каждой задаче имеется связка с соответствующим запросом в поисковик (связь один к одному). Задача на парсинг ставится на основе данных контента.

В Системе реализован процесс, который с периодичностью 15 минут инициирует создание задач. Для каждого имеющегося контента создаются задачи. Для каждой задачи формируется поисковой запрос, который будет использован при парсинге. Запрос формируется исходя из типа парсинга:

* Общий (имя контента + один из суффиксов (пример "Терминатор 2022 смотреть онлайн"));

В системе есть 2 типа суффиксов:

* + Суффиксы к конкретному контенту (задаются в запросе POST /films);
  + Общие суффиксы (задаются запросом POST /suffixes) — применяются ко всему контента.

Если к контенту не относится ни один суффикс, то запросы с этим контентом создаются без суффикса.

* Точный (site:sitename и filmname (пример "site:lordfilm.allcheck.online патриот 2020")).

2.4 Анализ html кода страницы

Система Content Security, в результате автоматизированного сбора и структурирования информации по поисковому запросу, получает список URL сайтов, которые необходимо проанализировать по требованиям пользователя.

Первоначальная проверка:

* Система проверяет наличие названия контента, который содержится в поисковом запросе, в title сайта;
* Система проверяет наличие года выпуска в description сайта или title сайта;

В случае, если данные условия выполняются, то происходит дальнейшая проверка:

* Система анализирует html код страницы, происходит поиск необходимых тегов/блоков кода, которые свидетельствуют о наличии искомого контента.
* Также в процессе анализа html кода страницы происходит поиск имен актера и режиссера.

Для каждого сайта введен рейтинг. Рейтинг показывает релевантность сайта требованиям пользователя. Соответственно, чем больше рейтинг для сайта, тем более вероятно на нем будет размещен нужный и искомый контент.

Рейтинг формируется исходя из анализа html кода. Определенные фрагменты html кода имеют свой вес, который суммируется в итоговом рейтинге.

Помимо рейтинга каждый сайт имеет величину под названием приоритет. Приоритет - порядок выдачи данного сайта по данному поисковому запросу в результатах.

2.5 Загрузка и процесс работы со списком сайтов, являющихся доверенными (whitelist url)

В систему загружается список из url сайтов, которые заведомо не должны проверяться на наличие контента. Список данных сайтов обозначен как whitelist url (WU). Если сайт, находящийся в WU, попал в результат выдачи, то он не проверяется на наличие контента (не анализирует title, description и html код сайта).

2.6 Выгрузка итогового списка во внешнюю систему

Результаты парсинга сохраняются в отдельную таблицу parsingresult.

Данные содержат:

* Наименование контента и его характеристики (год, актер, режиссер);
* url;
* Рейтинг сайта;
* Приоритет;
* Результаты парсинга периодически выгружаются во внешнюю систему.