УДК 004.021

Орлов А. С., Севрюков С. Ю.

Реализация алгоритма поиска по субтитрам YouTube

1. Введение. В наши дни платформа хранения, доставки и показа видео YouTube пользуется огромной популярностью. Она обладает большим спектром возможностей, что позволяет упрощать процессы работы с видеоматериалами. Одним из таких процессов является поиск по тексту субтитров соответствующих YouTube видео, генерируемых автоматически, что предоставляет возможность реализации поиска. Это позволит повысить производительность организаций, активно использующих YouTube в работе.

Примером такой организации может быть Центр развития электронных образовательных ресурсов (ЦРЭОР) Санкт-Петербургского государственного университета, который занимается созданием и выпуском онлайн-курсов. В процессе их создания видеолекции просматриваются авторами и сотрудниками ЦРЭОР для поиска ошибок. В курсах для публикации на платформе Coursera в среднем 4–6 модулей; на платформе Открытое Образование — 10–12 модулей. Каждый модуль включает в себя около 1,5 часа видеолекций. На один курс приходится минимум 6 часов итогового видеоматериала, не считая чернового. В практике центра часто встречались следующие ситуации: автор допускает возможность наличия терминологической ошибки при записи лекции; студенты указывают на ошибку в лекции, но не указывают на конкретный момент видео; автор хочет найти модуль и урок, в котором рассматривается определённая тема и т.п. Подобные задачи решаются вручную, что отрицательно сказывается на продуктивности при работе с видео контентом. Для повышения продуктивности, скорости и точности поиска фрагментов видео предлагается разработанный для этого алгоритм и прототип его реализации, который является частью разрабатываемого прототипа системы управления ЦРЭОР.

Орлов Антон Сергеевич – студент, Санкт-Петербургский государственный университет; e-mail: ion6431@gmail.com, тел.: +7(999)229-50-25

Севрюков Сергей Юрьевич – старший преподаватель, Санкт-Петербургский государственный университет; e-mail: s.sevryukov@spbu.ru, тел.: +7(911)910-94-88

2. Постановка задачи. В процессе исследования проблемы были определены входные ему данные:

1. Массив видео, опубликованных на YouTube, по которым будет производиться поиск.
2. Данные учётной записи Google.
3. Поисковый запрос.

Необходимо подчеркнуть, что качество генерируемых субтитров напрямую зависит от качества звука в видео [1]. Описываемая реализация направлена на повышение эффективности работы с видео в формате, предустановленном ЦРЭОР, однако может быть применена и к любым другим видео, удовлетворяющих начальным условиям.

Цель заключается в упрощении поиска, повышении его скорости и точности. Задача: реализовать алгоритм в виде прототипа. Реализуемая функциональность пользователя состоит из интерфейса индексации и интерфейса поиска, который по поисковому запросу возвращает список ссылок на на позиции в видео, отсортированные в порядке уменьшения релевантности.

3. Сущсествующие альтернативы. Было найдено расширение для браузера Google Chrome позволяющее искать по субтитрам YouTube [2]. Оно позволяет найти подстроки в субтитрах и указать моменты начала этих подстрок. Выполняется такой поиск только для одного видео. В реализованном прототипе выполняется поиск фраз, которые не обязательно должны быть подстроками субтитров, а область поиска не ограничивается одним видео.

4. Получение текста видео. Извлечение текста было произведено с помощью YouTube Data API. В него входит метод `captions.download`. Согласно документации, метод требует авторизации, следовательно, её необходимо реализовать.

4.1. Авторизация. YouTube API поддерживает два способа авторизации: OAuth 2.0 и API Key. Второй способ не предусматривает получение прав, поэтому в дальнейшем он рассматриваться не будет [3]. Для использования метода необходимо наличие хотя бы одного из двух разрешений: `youtubepartner` или `youtube.force-ssl`. Они указываются в параметре `scope`. Оставшиеся действия описаны в руководстве API [4]. После прохождения авторизации становится возможным использование метода получения субтитров.

4.2. Получение субтитров. Метод `captions.download` принимает в качестве обязательного параметра `id` дорожки субтитров. Его можно получить с помощью метода `captions.list`. Google накладывает условие, согласно которому субтитры возможно получить только из видео, загруженных пользователем. Данное условие может быть снято, если пользователь разрешает взаимодействие третьим лицам с его видео. Организациям, которым требуется поиск по собственным видео, это условие не создаст препятствий.

5. Индексация и поиск. Чтобы была возможность осуществить поиск, необходимо организовать хранение субтитров. Если просто извлечь их с помощью методов, описанных выше, то для каждого поискового запроса придётся выполнять обработку текста снова. Для ускорения этого процесса субтитры будут храниться уже в обработанном виде. В качестве поискового движка был выбран Elastic Search, в который встроены функции индексации и поиска [5].

Перед тем, как индексировать субтитры, необходимо определиться, что будет выступать в роли документа.

5.1. Заголовок как документ. Субтитры, генерируемые с помощью YouTube, являются множством заголовков, разделённых пустыми строками. Заголовок представляет из себя две строки: текст, полученный в результате распознавания речи, и временной интервал (примерно 2–5 секунд), в который эта речь прозвучала.

Если каждый документ будет представлять из себя заголовок с указанием времени, то результатом поиска будет позиция в видео, где в течении нескольких секунд прозвучит слово из запроса. Однако стандартное разбиение YouTube может относить слова из фразы в разные заголовки, что снизит качество поиска.

5.2. Видео как документ. Принимая объединение всех заголовков видео в качестве документа решается проблема разделения фраз, но результат поиска не будет указывать на позицию в видео.

5.3. Фрагмент как документ. Временные промежутки соседних заголовков могут пересекаться (тогда во время просмотра этого видео с субтитрами пользователь видит сразу несколько строчек субтитров). Однако если существуют моменты видео, в которых нет речи, в этом временном промежутке не генерируются субтитры. Следовательно, когда делаются паузы в речи, субтитры тоже не генерируются. Такие паузы можно вычислить, сравнивая конец

временного промежутка с началом следующего. Если второй начинается раньше, чем кончается первый, то есть пересечение, и паузы в речи не было. В противном случае пауза была, и это может значить, что сменилась тема разговора. По таким паузам происходит деление файла субтитров на фрагменты. Такое деление не разделяет слова во фразах, а длительность документов становится меньше или равной длительности видео целиком. Для наилучших результатов необходимо наличие пауз длиной 2-3 секунды в речи.

Способ дробления выбирается исходя из формата видео. Если предполагается, что поисковые запросы будут состоять из одного слова, то вариант дробления "заголовок как документ" будет выдавать наиболее точные результаты поиска. Когда массив видео состоит из коротких видео, внутри которых легко ориентироваться, то в роли документа лучше взять видео целиком. Для видео с долгими паузами в речи третий вариант будет наилучшим из перечисленных. Для ЦРЭОР последний вариант подходит лучше всего, потому что формат лекций предполагает наличие пауз. Разделив видео на документы по выбранному принципу, можно провести индексацию с помощью Elastic Search. Функция поиска тоже реализована в этом движке.

.

6. Тесты. В таблице 1 приведена информация о затратах времени для индексации при разбиении документ = заголовок

Таблица 1. Время индексации

Длительность видео	**Время индексирования**
7:52	0:18
1:52:36	1:18
2:55:02	5:04
35:34	2:02
53:19	3:46

При использовании двух других упомянутых способов разбиения субтитров время индексации заняло не больше трёх секунд. Время поиска не превышает одной секунды.

7. Вывод. Для повышения эффективности данный результат может быть интегрирован в информационную систему, в которой присутствуют компоненты взаимодействия с материалами, опубликованными на YouTube. Рассмотренные вариации методов поиска

показывают, что они могут варьироваться в зависимости от формата видео, для которых необходимо внедрение поиска. Алгоритм авторизации может быть модифицирован таким образом, чтобы пользователю не нужно было проходить авторизацию самостоятельно. В ЦРЭОР прототип был апробирован. Общий результат показал, что поиск с использованием прототипа повышает скорость и точность поиска, что доказывает достижение цели этой работы.

Литература

1. Как создать субтитры автоматически — Справка — YouTube [Электронный ресурс]: URL:https://support.google.com/youtube/answer/6373554?hl=ru (дата обращения: 14.03.18).

2. Текстовый поиск по видео на YouTube [Электронный ресурс]: URL:https://habrahabr.ru/post/321022/ (дата обращения 01.04.18).

3. YouTube Data API Overview | YouTube Data API | Google Developers [электронный ресурс]: URL:https://developers.google.com/youtube/v3/getting-started (дата обращения 01.04.18).

4. Implementing OAuth 2.0 Authorization | YouTube API | Google Developers [Электронный ресурс]: URL:https://developers.google.com/youtube/v3/guides/authentication (дата обращения 14.03.18).

5. Open Source Search and Analytics · Elasticsearch | Elastic [Электронный ресурс]: URL:https://www.elastic.co/ (дата обращения: 17.03.18).