УДК 004.89

ИСПОЛЬЗОВАНИЕ ВЕКТОРНЫХ ПРОСТРАНСТВ ДЛЯ РЕАЛИЗАЦИИ ТЕКСТОВОГО ПОИСКА ПО БОЛЬШИМ ОБЪЕМАМ МЕДИА ДАННЫХ

Петров О.Е. (Университет ИТМО)
Кабаров В.И. (Университет ИТМО)
Научный руководитель – д.т.н. Матвеев Ю.Н.
(Университет ИТМО)

В работе описывается механизм индексации большого объема медиа данных с использованием векторных пространств. Такая структура позволяет осуществлять быстрый поиск ключевых слов в аудио записях, используя текстовые запросы в свободной форме.

Благодаря техническому прогрессу и развитию сети Интернет объемы доступной для обработки информации кратно увеличивается год от года. Алгоритмы и методы поиска информации по тексту стали активно применятся несколько десятков лет назад вместе с ростом популярности сетевых поисковых сервисов, таких как Google. Вместе с этим начали развиваться алгоритмы поиска информации и по медиа данным. Современные технологии автоматического распознавания речи позволяют осуществлять поиск в том числе и по фразам, произнесенным в аудио записи. Целый ряд алгоритмов решает задачу поиска ключевых слов в фонограммах, содержащих речь. Предложенный в настоящей работе алгоритм позволяет строить индекс по результатам распознавания речи.

Процесс подготовки индекса для поиска включает в себя несколько этапов. На первом шаге с помощью системы автоматического распознавания речи выполняется преобразование речи в формат, включающий информацию не только о лучшей гипотезе распознавания, но и менее вероятных. В работе в качестве структуры данных используется сети спутывания. Полученные в процессе распознавания сети спутывания преобразовываются к векторам в заранее подготовленном векторном пространстве и размещается в индексе, который может включать в себя не всю фонограмму, а только некоторые ее части.

Для поиска информации текст запроса приводится к тому же векторному пространству, после чего выделяются самые похожие документы на основе косинусного сходства векторов запроса и каждого из документов индекса.

В рамках исследования был создан прототип, реализующий предложенный метод и получены базовые показатели качества работы системы.