УДК 004.822

МЕТОДЫ РАСШИРЕНИЯ ПОИСКОВЫХ ЗАПРОСОВ ПРИ МЕНЕДЖМЕНТЕ МЕДИА АКТИВОВ

Дубовской А.А. (Национальный исследовательский университет ИТМО)
Научный руководитель – доцент, к.т.н. Баймуратов И.Р.
(Национальный исследовательский университет ИТМО)

Введение. В силу неоднозначности естественного языка запросы в поисковых системах часто возвращают большое количество нерелевантных результатов. Чтобы увеличить количество релевантных документов, в запросах необходимо устранять неоднозначность, рассматривая их контекст. Для разрешения неоднозначности используются методы расширения запросов, такие как механизмы обратной связи по релевантности и онтологии. Существуют различные подходы к расширению запроса. Например, в статьях [1] и [2] представлены подходы к анализу знаний, поддерживающие расширение запросов, основанный на ассоциативных правилах, а в работе [3] авторы сочетают ассоциативные правила с моделью ранжирования. В этой работе рассматривается подход к интерактивному расширению запросов, основанный на ассоциативных правилах, областью его применения является управление медиа активами.

Основная часть. Для хранения ассоциативных правил был сформирован поисковый индекс в сервисе Elasticsearch. Формирование индекса включало следующие этапы:

1) Предварительная обработка метаданных медиа активов;
2) Формирование индекса медиа активов;
3) Формирование индекса тегов медиа активов;
4) Поиск ассоциативных правил между тегами;
5) Запись ассоциаций в индекс тегов.

Алгоритм процесса интерактивного расширения запроса состоит из следующих шагов:

1) Поисковая строка, введенная пользователем, фильтруется с помощью регулярного выражения и разбивается на отдельные слова, каждое из которых лемматизируется. Дополнительно из полученного списка удаляются стоп слова.
2) Отправляется запрос к индексу тегов elasticsearch, выполняющий поиск по точному совпадению токена со значением существующего тега и возвращающий его id. Генерируется список обязательных тегов запроса.
3) Отправляется запрос к индексу тегов elasticsearch, выполняющий поиск по полученным id из предыдущего пункта и возвращающий список ассоциаций тега.
4) Вывод полученных токенов в интерфейсе. Генерация расширенного списка тегов, выбранных пользователем.
5) Поиск в индексе медиа активов с использованием обязательного и расширенного списка тегов и вывод релевантных результатов запроса.

В качестве пользовательского интерфейса был использован сервис Streamlit.

Выводы. В данном исследовании был предложен интерактивный метод расширения запросов для управления медиа активами. Подход основан на поиске ассоциативных правил между тегами медиа активов, использует elastic search в качестве поисковой системы. В дальнейшем планируется дополнить систему управления медиа активами другими методами расширения запросов.

Список использованных источников:

1. Latiri C., Haddad H., Hamrouni T. Towards an effective automatic query expansion process using an association rule mining approach // Journal of Intelligent Information Systems 39. – 2012. – P. 209–247.

2. Liu C., Qi R., Liu Q. Query expansion terms based on positive and negative association rules // 2013 IEEE Third international conference on information science and technology (ICIST). – 2013. – P. 802–808.

3. Bouziri A., Latiri C., Gaussier E. LTR-expand: query expansion model based on learning to rank association rules // Journal of Intelligent Information Systems 55. – 2020. – P. 261–286.