РАЗРАБОТКА API-СЕРВИСА И ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ ДЛЯ ВСТРАИВАНИЯ ПОИСКА НА САЙТАХ

Сидоров С.В., Черкасова В.А.

Астраханский государственный университет им. В. Н. Татищева,
Россия, 414056, Астрахань, ул. Татищева 20а
тел.: +79378296885, email: galserge.sidorov@gmail.com
Астраханский государственный университет, Россия, 414056, Астрахань, ул. Татищева 20а, тел.: +79275886081, email: valyc@mail.ru

В работе представлена разработка поисковой системы для официального портала Астраханского государственного университета https://asu.edu.ru/. Актуальность работы обусловлена необходимостью оперативного получения информации пользователями в условиях существования больших объемов данных на сайте. Используемый ранее встраиваемый поиск на сайте от Яндекса не удовлетворял в плане качества и возможности настройки поисковой выдачи. Программные решения для полнотекстового поиска вроде Elasticsearch и Sphinx имеют недостаточно гибкий интерфейс, что не позволяет внедрить их поиск на сайте ввиду особенностей организации данных. Поэтому возникла необходимость в разработке поисковой системы, удовлетворяющей потребностям портала.

Поиск осуществляется на основе индексирования, что позволяет ускорить процесс оценки и извлечения информации, соответствующей запросу пользователя [1].

Индексирование данных может быть реализовано различными методами [1]. Наиболее распространенным является, так называемый, инвертированный индекс, но при использовании данного подхода возникают сложности с ранжированием документов. Результатом данного проекта стала реализация поискового API, использующего TF-IDF для представления связей между документами и терминами в них [2].

Поисковый API разработан на базе фреймворка fastAPI, его прототип используется на сайтах Астраханского государственного университета. Благодаря тому, что поиск реализован как API-сервис, он может обслуживать несколько сайтов, расположенных на разных серверах. При этом процесс индексирования осуществляется непосредственно через базы данных. Таким образом, разработанное решение объединяет в себе особенности двух наиболее распространенных систем Elasticsearch и Sphinx. В дальнейшем планируется доработка интерфейса API (упрощение принципа работы) и реализация функций автодополнения, реферирования текста и исправления ошибок и опечаток в запросах.

Литература

1. *Марманис Х, Бабенко Д.* Алгоритмы интеллектуального интернета. Передовые методики сбора и обработки данных. – СПб.:Символ-Плюс, 2011. – 480 стр.
2. Speech and Language Processing (3rd ed. draft) [Электронный ресурс] / Dan Jurafsky, James H. Martin. – Режим доступа : https://web.stanford.edu/~jurafsky/slp3/, свободный.