

Введение. Сервисы объявлений являются одними из самых популярных интернет-ресурсов на сегодняшний день. Основной функционал таких сервисов заключается в предоставлении листинга – списка документов, соответствующих заданным пользователем условиям. Подобные системы можно классифицировать как системы атрибутивного поиска документов. Такие поисковые системы обычно обладают набором заранее определённых фильтров, количество которых увеличивается с ростом числа атрибутов документов, что создаёт когнитивную нагрузку на пользователей. Это вызывает необходимость в разработке упрощённой системы пользовательского ввода фильтров, способной преобразовывать запросы на естественном языке в структурированные запросы к нижележащей поисковой системе. В работе [1] рассмотрена данная проблема, решение которой будет полезно как для обычных пользователей сервисов, так и для специалистов, которым необходимо взаимодействовать с поисковой системой без знания её структуры и форматов запросов.

Основная часть. В ходе исследования были рассмотрены и проанализированы различные подходы к решению поставленной задачи. Задача текстового информационного поиска может быть решена с использованием вероятностных моделей ранжирования, таких как BM25, которая была рассмотрена в работе [2]. Однако для решения основной задачи преобразования запросов на естественном языке в поисковые запросы можно опираться на тот факт, что типовые запросы имеют схожую структуру. Например, для сервиса продажи автомобилей запрос может выглядеть следующим образом: «Красная Лада Веста от 1 миллиона до 2 миллионов рублей с пробегом». Для распознавания семантики запроса достаточно построить формальную грамматику языка запросов.

Таким образом, возникает необходимость в создании доменно-специфичного языка (DSL), который позволит задавать грамматику и генерировать парсеры для распознавания запросов на естественном языке. Были рассмотрены варианты использования внешних DSL, таких как YACC или ANTLR. Однако, помимо недостатков, присущих внешним DSL, которые были рассмотрены в работе [3], они оказались слишком универсальными для решения поставленной задачи. Ключевые недостатки заключаются в отсутствии поддержки динамической генерации правил на основе словарей, которые обновляются с определённой периодичностью, а также в сложности их интеграции в существующие проекты.

В связи с этим было принято решение разработать собственный внутренний DSL для описания грамматики запросов на языке Scala. Язык Scala был выбран по причине его гибкости синтаксиса, что делает его одним из наиболее подходящих языков для создания внутренних DSL. Кроме того, благодаря функционированию на платформе JVM, его легко интегрировать в сервисы, разработанные на более популярных языках, таких как Java.

Разработанный DSL включает набор инструментов для описания грамматики и типовых структур (например, диапазоны «от X до X», числовые литералы). Основное преимущество предложенного решения заключается в поддержке динамического обновления словарей и генерации новых правил, что особенно полезно для работы с каталогами, например, автомобилей, где марки и модели могут регулярно обновляться. Таким образом,

разработанный DSL представляет собой универсальное средство для задания доменно-специфичной грамматики поисковых запросов.

В заключение была рассмотрена возможность расширения функциональности поисковой системы для реализации поиска по другим разделам сервиса, таким как статьи или отзывы пользователей. На основе идей, изложенных в работе [4], было предложено решение, сочетающее использование вероятностных моделей ранжирования с векторными моделями.

Выводы. В рамках данной работы предложен подход к реализации DSL для задания грамматики и генерации парсеров, способных распознавать типовые поисковые запросы и преобразовывать их в структурированные запросы к основному поисковому сервису. Разработанный DSL был реализован в виде прототипа на языке Scala и протестирован на популярном сервисе объявлений Auto.ru.

Список использованных источников:

1. Посевкин Р.В., Бессмертный И.А. Translation of natural language queries to structured data sources // 9th International Conference on Application of Information and Communication Technologies (AICT), Rostov on Don, Russia – 2015 – С. 57-59.
2. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. // Foundations and Trends in Information Retrieval – 2009 – С. 333-389.
3. Shen, L., Chen, X., Liu, R. Domain-Specific Language Techniques for Visual Computing: A Comprehensive Study. // Arch Computat Methods Eng 28 – 2021 – С. 3113–3134
4. Nogueira R., Yang W., Cho K., Lin J. Multi-Stage Document Ranking with BERT // arXiv URL: <https://arxiv.org/abs/1910.14424> (дата обращения: 23.02.2025).

Автор _____ Цю Т.
Научный руководитель _____ Саржевский И.А.