УДК 004.82

МОДУЛЬ ИМПОРТА ОБЪЯВЛЕНИЙ ОБ ОБЪЕКТАХ НЕДВИЖИМОСТИ ИЗ ВНЕШНИХ ИСТОЧНИКОВ В СИСТЕМУ ХРАНЕНИЯ НА БАЗЕ СЕМАНТИЧЕСКОЙ СЕТИ

Нестеров И.А. (Университет ИТМО)
Научный руководитель — Цопа Е.А. (Университет ИТМО)

Введение. В эпоху современных технологий важность когнитивного восприятия информации из информационных систем играет особую роль. Наиболее естественной же для человека формой восприятия информации является организация данных в виде семантической сети. Так устроен человеческий мозг. Наиболее же популярные информационные системы, хранящие объявления об объектах недвижимости, хранят информацию иначе: в виде набора несвязанных друг с другом характеристик объекта, сами объявления зачастую представлены в виде списка. Подобный формат хранения не позволяет использовать достоинства семантических сетей, будь то решение задач выборки информации из текстов, написанных на естественном языке или же использование запросов к некой базе данных с использованием композиционных и родо-видовых связей. Реализация импорта объявлений в систему хранения на базе семантической сети, разрабатываемой в Университете ИТМО [1], позволяет использовать вышеописанные возможности работы с данными.

Основная часть. В качестве внешних источников было принято решение взять популярные российские интернет-сервисы для размещения объявлений о недвижимости: «Циан» и «Авито». В HTML-странице объявления в этих источниках собрана репрезентативная информация об объекте недвижимости: площадь, количество комнат, год сдачи, местоположение и прочее. Была сформирована выборка из нескольких сотен объявлений с различным наполнением из обоих источников, которые были проанализированы с целью определить наиболее репрезентативную и при этом универсальную структуру объявления. В дальнейшем было принято решение модифицировать получаемые объявления в сформированную по результатам этого исследования структуру данных. Структура была описана в существующей семантической сети.

При помощи экосистемы фреймворка Spring был реализован механизм получения содержимого HTML-страниц в разработанное приложение для его дальнейшего анализа. Средствами языка программирования Java был реализован механизм парсинга HTML-страниц, основанный на DOM-навигации и регулярных выражениях. Далее полученная информация реорганизуется в иерархическую структуру: определяются виды отношений и семантические связи между концептами [2]. Объявление преобразуется в вышеописанную структуру. После этого преобразованное объявление сохраняется в базу данных. В системе обновляются связи между элементами данных, добавляются связки с недавно добавленными.

В ходе разработки приложения возникла сложность, заключающаяся в необходимости по-разному обрабатывать HTML-страницы из разных внешних источников, поскольку каждый интернет-сервис демонстрирует разные структуры HTML-страниц. В связи с этим возникла необходимость реализовать специфичный алгоритм под каждый внешний источник объявлений. Однако для упрощения импорта объявлений пользовательский интерфейс был унифицирован. Приложение было спроектировано по архитектуре REST и ожидает на вход гиперссылку на объект недвижимости, приложение вычленяет из нее источник и верифицирует его; далее, в случае успешной проверки, перенаправляет дальнейшую обработку на соответствующий сервис.

По итогам реализации было проведено функциональное и нагрузочное тестирование, позволившее оценить эффективность работы приложения по импорту объявлений, в том числе при условии работы под нагрузкой.

Выводы. В рамках работы было реализовано приложение по импорту объявлений об объектах недвижимости из внешних источников («Циан» и «Авито»). Программа была реализована в виде Spring-Boot приложения и протестирована на примере извлечения информации из реально существующих объявлений из вышеописанных внешних источников, показав свою эффективность.

Список использованных источников:

1. Клименков С.В., Николаев В.В., Харитонова А.Е., Гаврилов А.В., Письмак А.Е., Покид А.В. Применение семантической сети для хранения слабоструктурированных данных // Инженерный вестник Дона [электронный журнал] – 2020. - № 2(62). – С. 27
2. Письмак А.Е., Харитонова А.Е., Цопа Е.А., Клименков С.В. Метод автоматического формирования семантической сети из слабоструктурированных источников. Программные продукты и системы. 2016. № 3. С. 74–78.