УДК 004.02

МЕТОД ИСПОЛЬЗОВАНИЯ ФЕДЕРАТИВНЫХ SPARQL ЗАПРОСОВ ДЛЯ ДОСТУПА К УДАЛЕННЫМ ИСТОЧНИКАМ ОТКРЫТЫХ ДАННЫХ

Волков А.А. (Университет ИТМО)
Научный руководитель – к.т.н, Тесля Н.Н.
(Университет ИТМО)

В работе рассмотрен вопрос расширения информации из открытых данных карточек ДТП дополнительными историческими данными о погоде. Данные о погоде предоставляются внешними веб сервисами, которые имеют ограничение на количество запросов в сутки и формат выходных данных. Для объединения данных возможно использования федеративного SPARQL, который будет возвращать пользователю расширенный граф знаний.

Введение. Результаты анализа данных напрямую связаны с качеством источников данных. В данных могут встречаться пропуски или некорректные значения. При агрегации данных чаще всего ставится одна задача в контексте которой данные и будут в дальнейшем использованы. С учетом этого осуществляется оптимизация хранения и доступа информации, с целью повысить быстродействие и уменьшить стоимость решения. Такая фокусировка является проблемой при решении междисциплинарных, комплексных, задач, так как в процессе работы требуются объединять несколько источников данных, которые представлены в разных форматах. Для решения данной проблемы существуют разные решения по объединению данных: Data Lake, Data Warehouse, Data Fabric. Эти решения подходят в случае, если все используемые источники данных локальные. Для комплексного решения, когда часть источников данных локальные, а часть удаленные такие способы объединения не применимы.

Основная часть. Исходной задачей является расширение информации, представленной в карточках ДТП, дополнительными данными о погоде. В карточках ДТП нет подробной информации о погодных условиях, только краткое словесное описание освещенности и осадков. В основном это ограниченный набор возможных значений. В некоторых случаях информация о погоде противоречива, например, указания сильного тумана и полной видимости. Поэтому в данной работе предлагается использовать сторонние источники данных о погоде, которые имеют ограничение на количество запросов в сутки. Рассмотрим алгоритм работы метода. Пользователь платформы формирует SPARQL запрос, в котором указываются необходимые данные из SPARQL endpoint и флаг расширения данных. Далее запрос отправляется в сервис карточек ДТП, который представляет собой SPARQL endpoint. Сервис формирует SQL запрос в базу данных карточек используя R2RML маппинг. После получения ответа из базы данных, сервис карточек ДТП отправляет результат запроса в формате RDF в ядро, которое отправляет исходный запрос SPARQL и данные полученные на прошлом шаге в сервис погоды. Сервис погоды опрашивает кеширующую базу данных о наличии данных, которые расширяют RDF. В случае отсутствия необходимых данных сервис погоды запрашивает данные в веб сервисе в формате HTTP. После этого веб сервис отправляет результаты сервису погоды в JSON формате. Сервис погоды формирует и отправляет расширенные данные RDF ядру, которое возвращает пользователю данные в RDF.

Выводы. Объединение нескольких источников данных позволяет повысить качество выходных данных и упростить дальнейший анализ. В случае, когда это объединение происходит незаметно для пользователя это повышает удобство использования источников данных.