Авторы:
Н.А. Рогаленко, А.С. Яркеев, А.Д. Пашнин, Университет ИТМО, Санкт-Петербург
Научный руководитель:
Цопа Евгений Алексеевич, Университет ИТМО, Санкт-Петербург

Тезис доклада:

Наполнение семантической сети адресной информации

Одной из предметных областей, имеющих научный и практический интерес для представления в качестве семантической сети, является адресное пространство городов и населённых пунктов. Использование данной сети позволяет более эффективно управлять адресными данными в прикладных программах, а также предоставляет широкие возможности для манипуляции данными с целью автоматизации процессов, связанных с поиском, определением и выборкой адресов. Целью данной работы является наполнение семантической сети адресной информацией.

В ходе анализа возможных источников для наполнения сети была выбрана Федеральная информационная адресная служба (сокращённо ФИАС). Это обусловлено рядом преимуществ, предоставляемых данной системой, таких как достоверность всей информации, регулярное обновление, структурированность, наличие возможности однозначно идентифицировать каждый объект и оценить актуальность данных о том или ином адресном объекте. На основе данных ФИАС была получена наиболее полная информация о муниципальном делении, территориальной и административной принадлежности объектов. Формат данных ФИАС позволяет однозначно определять актуальность наименований объектов и производить обновление информации без полного пересоздания системы.

Для создания семантической сети и её заполнения данными была разработана программа, обрабатывающая информацию из ФИАС и генерирующую на её основе скрипт на специальном декларативном языке SemQL, используемом для работы с семантическими сетями. После создания сгенерированный скрипт обрабатывался модулем импорта данных в семантическую сеть.

В процессе данной работы был разработан набор скриптов на языке программирования Perl, предназначенный для заполнения семантической сети адресными данными, обрабатывающий входной XML файл из базы ФИАС и создающий на его основе SemQL скрипт. Особое внимание при разработке уделялось выборке только актуальных адресов, поскольку устаревшие наименования не удаляются из базы, а получают соответствующие значения атрибутов. В скрипте, являющимся результатом работы программы, содержится код для генерации всех необходимых смысловых значений сети и их экземпляров, а также возможные вордформы для некоторых объектов, для поиска которых также был создан отдельный алгоритм и определенный набор правил.

Результат проделанной исследовательской работы позволяет автоматически получать самую полную и актуальную информацию об адресных объектах и заполнять полученными адресными данными семантическую сеть. В конечном итоге, на основе результатов работы сгенерированного SemQL, была создана семантическая сеть, содержащая более миллиона экземпляров и вордформ и более ста пятидесяти смысловых значений.