УДК: 004.624

Импорт именованных сущностей из реляционной БД в семантическую сеть

А.О. Ванцев

(Университет ИТМО, Санкт-Петербург)

Научный руководитель: старший преподаватель Е.А. Цопа

(Университет ИТМО, Санкт-Петербург)

В работе рассматривается способ создания семантической сети на основе метаданных реляционной базы данных и импорта из последней именованных сущностей. Автором разработано приложение, результатом работы которого является скрипт на языке SemQL.

Введение

Анализ текстов на естественном языке является актуальным и перспективным направлением в области автоматической обработки текстов и компьютерной лингвистики. Исследования в этой области носят как научный, так и практический интерес, поскольку существует множество задач, которые решаются с помощью вычислительных устройств. Эти задачи можно классифицировать по уровню детализации: сигнала, слова, словосочетаний, предложений и т.д. На уровне словосочетаний происходит определение частей речи, выделение слов, распознавание именованных сущностей. Для решения последней задачи требуется иметь семантическую сеть, не только определяющую модель предметной области, но и содержащую конкретные экземпляры.

Цель

Целью данной работы является автоматизация создания семантической сети на основе метаданных реляционной базы данных и последующего импорта из последней именованных сущностей.

Базовые положения исследования

Создание семантической сети требует информации о предметной области: ее объектах и связях. Вследствие этого автором был предложен следующий подход к формированию сети:

1. Анализируется структура метаданных реляционной БД: таблицы, колонки, связи между таблицами.
2. На основе результатов анализа метаданных реляционной БД формируется структура семантической сети (узлы и семантические отношения), описывающей заданную предметную область.
3. На основе созданной модели формируется определенная выборка из реляционной БД, позволяющая наполнить семантическую сеть уникальными именованными сущностями.

Для проверки применимости предложенного подхода к решению практических задач автором было разработано приложение, реализующее следующие функции:

1. Подключение к реляционной БД и получение ее метаданных.
2. Предоставление графического интерфейса для отображения обнаруженных таблиц и связей между ними в виде диаграммы.

3. Автоматическое создание семантической сети на основе выбранных пользователем таблиц и колонок.
4. Получение из реляционной БД именованных сущностей.
5. Создание и сохранение в текстовый файл скрипта для создания желаемой семантической сети и наполнения ее данными

Результаты

Приложение было выполнено на языке программирования Java. Предусмотрена возможность подключения к различным БД: MS SQL, Oracle, PostgreSQL, однако архитектура приложения позволяет легко добавлять новые источники данных.

Список всех найденных таблиц отображается во вспомогательном окне, а отмеченные пользователем – в основном окне в виде ER-диаграммы, динамически изменяющейся при добавлении или удалении таблиц.

Структура семантической сети создается при выборе нужных таблиц или их колонок и отображается в отдельной вкладке.

При условии корректной структуры генерируется скрипт на языке SemQL.

Ванцев А.О. (автор)

Цопа Е.А. (научный руководитель)