УДК 004.822

Извлечение и структурирование данных о компьютерных системах

Редькина И.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – старший преподаватель Цопа Е.А.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В докладе рассматривается способ расширения семантической сети, которая была получена на основе составления универсальной модели компьютера, вместе с её заполнением и дальнейшим использованием при обработке неструктурированных текстов.

Введение. Существует множество различных видов и характеристик компьютера, которые могут идентифицировать его как уникальную сущность. Все эти составляющие важны при распознавании данных о компьютере из неструктурированного текста - поиск в интернет-магазине или сайте производителя. Поскольку производители составляющих компьютера создают идейно одинаковые компоненты, отличающиеся только техническими характеристиками, нужно обозначить структуру, которая позволяет работать с ними как с одной сущностью, имеющей только различные характеристики. И семантические сети, представляющие собой ориентированный граф, могут собрать все эти данные в такой вид, который можно было бы использовать для дальнейшей обработки, то есть появится возможность составить универсальную модель компьютера, которая не будет зависеть от каких-то конкретных параметров, а будет использована для получения структурированных данных из необработанных текстов.

Основная часть. Предлагается подход к обработке текста, который сможет распознавать неструктурированные данные, чаще всего являющиеся написанным человеком текстом, и наполнять созданную семантическую сеть. Для получения универсальной модели компьютера, которая и представляет модель семантической сети, были изучены все характеристики компьютера и его подвидов на сайтах интернет-магазинов с поиском по параметрам, сайты-производители составляющих компьютеров и другие источники, которые могут предоставить данную информацию. Для заполнения готовой семантической сети используются данные из открытых источников, которые предоставляют информацию с большим количеством реальных параметров для одного компьютера и конкретных значений. И после получения готовой для дальнейшей обработки модели разрабатывается сам алгоритм распознавания нужной информации из неструктурированных текстов для последующего заполнения семантической сети.

Выводы. В результате проведённой работы получена структура в семантической сети для универсальной модели компьютера, которая наполнена реальными обработанными данными и готова для дальнейшего использования. Такая готовая модель может использоваться при реализации автодополнения параметров компьютера при вводе пользователем данных при поиске или описании устройств. А полученный алгоритм распознавания текста будет обрабатывать больше текстов для дальнейшего расширения семантической сети.

Редькина И.В. (автор) Подпись

Цопа Е.А. (научный руководитель) Подпись