УДК: 004.822
Название: Сравнительное исследование методов восстановления связей в графе знаний социо-экономической системы

Авторы:
Калинин А.М, Университет ИТМО, г. Санкт-Петербург;
Боченина К.О., Университет ИТМО, г. Санкт-Петербург

Научный руководитель: Боченина К.О., Университет ИТМО, г. Санкт-Петербург

Тезис доклада:

Социо-экономическая система – это комплексная система, интегрирующая разные источники данных о поведении пользователей социальных систем, потребителей товаров и услуг финансовой сферы, компаний в сетях контрагентов и др. В данной работе такая система представляется в виде графа знаний, в который интегрируются три разных источника (социальная сеть, система интернет-рекрутинга, система хранения транзакционных данных физических лиц). В графе знаний все сущности системы представлены в виде узлов, а связи между ними – в виде рёбер, имеющих конкретный семантический смысл (отношение). Имея достаточно полную базу знаний, можно восстанавливать пропущенные (по разным причинам) связи между сущностями, тем самым, дополняя информацию и повышая полноту базы.

Целью исследования являются интеграция различных источников данных в единый граф знаний социо-экономической системы и проверка построенного графа на задаче восстановления связей между сущностями. Для создания единого графа знаний социо-экономической системы интегрированы три источника данных: социальная сеть «ВКонтакте», система интернет-рекрутмента «HeadHunter» и банк. В связи с тем, что данные и онтологии указанных источников пересекаются, необходимо проведение процедуры сопоставления, при этом схемы (онтологии), ввиду их небольшого размера, сопоставляются вручную. Данные о работодателях и профессиях в социальной сети «ВКонтакте» заполняются пользователями самостоятельно, без каких-либо ограничений. Как результат, написания одних и тех же организаций/профессий сильно различаются, что приводит к большому количеству дубликатов. Для решения данной проблемы разработан автоматизированный метод объединения схожих записей на основе предварительно обученной мульти-язычной модели отображения текстовых данных в векторные представления. Метод также применяется для сопоставления сущностей между источниками.

Конечная версия графа знаний содержит 11 типов сущностей, 14 отношений, 2,4 млн. сущностей и 128 млн. триплетов. Большая часть сущностей представляет пользователей и сообщества «ВКонтакте». Большая часть триплетов связывает указанные типы сущностей друг с другом. Для решения задачи восстановления связей выбраны методы на основе векторных представлений элементов графа знаний (TransE, ComplEx, RotatE) и графовая нейронная сеть HGT. В качестве эксперимента выбраны три отношения: Пользователь-Работодатель, Пользователь-Профессия, Пользователь-СЭС (социо-экономический статус). Результаты показали абсолютное преимущество метода на основе графовой нейронной сети с архитектурой Трансформер. Для 74% и 88% истинных триплетов Пользователь-Работодатель и Пользователь-Профессия, соответственно, правильный ответ ранжируется не хуже десятого места в общем рейтинге предполагаемых триплетов.

Результатом работы является конвейер по интеграции различных источников данных в единый граф знаний. Качество построенного графа проверено на задаче восстановления связей.

Автор _____ /_____Калинин А.М._____/

Научный руководитель _____ /_____Боченина К.О._____/