П. В. Юхневич

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕОБРАЗОВАНИЯ ПОИСКОВОГО ЗАПРОСА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Рассматривается информационный поиск. Предлагается использование моделей машинного обучения для упрощения формирования запроса к БД.

Введение

Для данной задачи предлагается использовать два последовательно применённых алгоритма: первый решает задачу классификации (для определения типа искомого объекта), второй задачу извлечения именованных сущностей (для установления соответствия искомых слов атрибутам соответствующего типа). Далее предлагается построить объект, с помощью которого осуществляется запрос к БД;

I. Описание процесса поиска

Имея запрос пользователя на естественном языке, необходимо понять, объект какого типа он пытается найти, то есть решить задачу классификации. Для этого предлагается использовать векторное представление.

Векторное представление (Word Vector Embedding) – сопоставление словам из словаря векторов из R^n для n, значительно меньшего количества слов в словаре [1].

Определённый тип искомой сущности содержит свой набор атрибутов и для того, чтобы определить соответствие атрибутов словам из поискового запроса на естественном языке, предлагается использовать определение именованных сущностей.

Определение именованных сущностей (Named Entity Recognition – NER) – получение структурированного представления информации по тексту на естественном языке, в нем содержащейся [2].

Для создания моделей необходима обучающая выборка, которую предлагается брать из БД, в которой хранится искомая информация в уже структурированном и упорядоченном виде: таблицы соответствуют типам, а поля и их значениям.

Результаты работы моделей предлагается использовать для формирования запроса к БД посредством ORM, используя оператор подобия.

II. Результат работы моделей машинного обучения

Для демонстрации предлагаемого подхода был проведён эксперимент, результаты которого (табл.1.) наглядно показывают работу упомянутых выше моделей. Была выбрана предметная область, содержащая 3 класса сущностей, общим объёмом 500 объектов, при этом обучение осуществлялось на 100 объектах.

Таблица 1 – Результат работы моделей

Запрос	Классификация	NER
Android 10	smartphone	os: Android, os_version: 10
notebook xiaomi	notebook	type: notebook, brand: xiaomi
xiaomi	smartphone	brand: xiaomi
apple invention	news	text: apple invention

Полученные результаты (например, результат классификации по запросу «xiaomi» и «xiaomi notebook») указывает на то, что в первом случае, определённый тип чаще других содержит данное ключевое слово, во втором же случае тип является другим, так как пользователь задал его явно в запросе.

Примечателен также и последний запрос, который мог бы быть отнесён к классу «smartphone», но был отнесён к классу «news», так как слово «invention» встречается там и не встречается в «smartphone».

III. Выводы

Предлагаемый подход позволяет определить тип искомой сущности и отобразить слова из поискового запроса на соответствующие атрибуты данного типа до обращения в базу данных, тем самым снизив число запросов.

1. Журавлев, Ю. И. Распознавание. Математические методы / Ю. И. Журавлев, В. В. Рязанов, О. В. Сенько // Информатика. – 2006. – №2. – С. 30-40.

Юхневич Павел Витальевич, магистрант кафедры интеллектуальных информационных технологий БГУИР, yukhnevichpavel@gmail.com.

Научный руководитель: Сердюков Роман Евгеньевич, доцент кафедры интеллектуальных информационных технологий БГУИР, кандидат технических наук, rserdyukov@gmail.com