Виртуальная лаборатория оценки качества векторных представлений объектов

Колосов Алексей Михайлович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: akolosov@cs.msu.ru

Научный руководитель *— Майсурадзе Арчил Ивериевич*

В настоящее время широко распространено применение векторных представлений объектов. Векторное представление это последовательность чисел, характеризующая «положение» объекта в некотором скрытом «пространстве». При этом похожие по смыслу объекты оказываются рядом в смысле заданной функции близости. Одним из способов использования векторных представлений является быстрый поиск ближайших соседей. Так, если в качестве объекта рассматривается последовательность слов, то, например, для запроса в поисковой системе возможно упорядочить выдачу документов по степени близости векторных представлений документов к векторному представлению запроса.

В работе в качестве объектов рассматриваются слова, поскольку для них есть размеченные экспертами наборы, содержащие сведения о семантической близости слов. Для оценки качества векторных представлений слов предлагается использовать виртуальную лабораторию[1] — программное средство, позволяющее определить основные показатели качества. Актуальность наличия такой лаборатории заключается в возможности сравнивать качество существующих векторных представлений.

Лаборатория позволяет автоматически оценить размерность представлений; проверить метричность для заданной функции близости на заданных представлениях; вычислить ранговую корреляцию с экспертным набором данных, используя представления как источник близостей.

Оценка качества производится с использованием экспертных наборов данных. В настоящее время лаборатория содержит наборы данных, которые состоят из пар слов и численных оценок близости. Дополнительно может содержаться информация о частях речи. Доступны следующие наборы экспертных данных: WordSim353, The MEN Test Collection, SimLex-999.

[1] https://github.com/obj2vec/validation_scheme, https://obj2vec.cs.msu.ru

В режиме оценки качества векторных представлений процесс работы лаборатории состоит из следующих шагов:

1. Поступает один из следующих вариантов входных данных:

   - Набор пар <объект, векторное представление> без функции близости — в этом случае выбирается одна из доступных функций расстояния: Евклидово, косинусное, Манхэттонское, Чебышёвское. По умолчанию используется косинусное расстояние.
   - Набор пар <объект, векторное представление> и программа с функцией близости.
   - Если векторных представлений нет или пользователь не готов ими делиться, возможен режим работы только с близостями — в этом случае на вход подаются тройки <объект1, объект2, величина близости>. Для слов это должны быть слова в начальной форме, без указания частей речи. Например, в качестве такого входа возможно подать свой экспертный набор данных.
   - Также возможен вариант входа без векторных представлений, когда на вход подается только набор объектов и программа, реализующая функцию близости. Этот вариант входа актуален для контекстно-зависимых представлений.

2. Выбираются экспертные наборы данных и показатели качества.

3. Вычисляются показатели качества для набора объектов, представленных и во входном, и в экспертном наборах.

Результатом работы лаборатории в режиме оценки качества является набор вычисленных показателей качества. При этом представления могут оцениваться не только для сравнения с другими представлениями, но и для того, чтобы стать новым экспертным набором.

Описанная лаборатория качества позволяет не только сравнивать качество векторных представлений, но и расширить множество наборов экспертных данных. Несмотря на то, что в настоящее время в лаборатории качества доступны только экспертные наборы со словами, в будущем будут добавлены и другие типы объектов.