А. М. Ковалевский

СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ В РЕШЕНИИ ЗАДАЧИ ПРОФИЛИРОВАНИЯ

Профилирование – разумное ограничение предъявляемой посетителю информации с целью выделения более важного для него содержания. Задачей профилирования является правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной пользователю информации [1]. Решение этой задачи позволит потребителям услуг тратить меньше времени на просмотр и усвоение контента и больше на ее практическое применение.

Введение

Задача профилирования пользователя представляет собой задачу классификации, т.е. определение к какому классу относится входной объект. Для решения поставленной задачи, решаем использовать сверточную нейронную сеть на основе кодирования слов с помощью семантической репрезентации.

I. Векторная репрезентация слов

Для обучения нейронной сети необходимо, представить в векторной форме с помощью алгоритма Word2Vec все интересующие тематики для пользователя. А также самих пользователей объединить в группы, при наличии схожих тематик. Пример использования алгоритма на рис.1.

Негативным эффектом векторной репрезентации является быстрая деградация векторов при операциях над ними. Сложение векторов двух слов демонстрирует то общее, что есть между этими словами, при условии что слова действительно связаны в реальном мире, но попытка увеличить количество слагаемых очень быстро приводит к потере какого-либо практически ценного результата [2]. Сложить слова одной фразы ещё выполнимо, сложить слова нескольких фраз уже нет. Необходим иной подход.

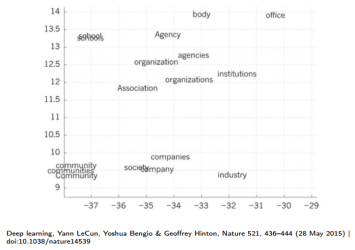

Рис. 1 – Пример работы алгоритма Word2Vec

II. Семантическая репрезентация слов и текстов

Из векторных репрезентаций слов создаётся семантический вектор смыслов слов. Для этого проведем кластеризацию вектора наших слов. А затем для каждого слова вычисляется расстояние до центра кластера и отбрасываются значения менее 0. Полученные расстояния до центра и есть искомый семантический вектор. Каждый элемент данного вектора имеет свой смысл, задаваемый теми словами, что образуют соответствующий кластер. Сложение таких векторов деградирует намного медленнее сложения оригинальных репрезентаций слов.

Складывая семантические вектора отдельных слов, составляющих текст, получается семантический вектор всего текста. Так как каждому тексту поставлен в соответствие вектор в семантическом пространстве, возможно вычислить расстояние между любыми двумя текстами как косинусную меру между ними. Имея расстояние между текстами, можно провести классификацию в векторном пространстве текстов, а не отдельных слов. Это необходимо для фильтрации самих текстов согласно требуемым тематикам профилирования.

III. Выводы

Представленный алгоритм определения профиля пользователя, основан на семантической репрезентации текста с использованием сверточной нейронной сети, что является абсолютно новым методом классификации пользователей. Анализируя просмотренные пользователем данные, мы можем спрогнозировать и предложить пользователю наиболее интересную для него информацию благодаря правильному кодированию интересующих пользователя слов.

Список литературы

1. Герман, О. В. Введение в теорию экспертных систем и обработку знаний / О. В. Герман // Минск, Дизайн-Про, 1995.
2. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский // Москва, Горячая Линия - Телеком, 2007.

Ковалевский Александр Михайлович, магистрант кафедры информационных технологий автоматизированных систем БГУИР, aliaksandr.kovalevsky@gmail.com.

Научный руководитель: Гуринович Алевтина Борисовна, кандидат физ.мат наук, доцент, gurinovich@bsuir.by.