УДК 004.032.26

СИСТЕМА ПРЕДИКТИВНОГО АНАЛИЗА ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

А.Л. Калоша
Магистрант. кафедры информатики БГУИР

М.А. Медунецкий
Профессор, доктор технических наук

М.П. Хоронеко
Студент БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: andreikalosha@mail.ru

Аннотация. Цель данной работы заключается в создании системы предиктивного анализа для классификации документов текстовых коллекций, например, публикаций. Классификация производится с использованием нейронной сети по набору метрик, описывающих качество публикаций. В настоящей работе в качестве набора метрик используется количество просмотров, лайков и репостов.

Ключевые слова: Big Data аналитика, TensorFlow, CUDA, машинное обучение, нейронные сети.

Скорость, с которой публикуется новая текстовая информация, растет с каждым днем. Изучить весь контент не представляется возможным даже в отдельных областях, поэтому приходится фильтровать получаемую информацию и выбирать лучшую. Назначение разрабатываемой системы заключается в предсказании популярности публикаций через заданный промежуток времени. Популярностью статьи считается количество просмотров, лайков или репостов, которые зависят от большого количества факторов, таких как время публикации, название, содержание статьи, авторов, и так далее. Данные параметры наилучшим образом отражают качество публикации. Использование машинного обучения позволяет предсказать значения метрик популярности для неопубликованного контента.

Для обучения нейронной сети была выбрана библиотека TensorFlow как один из лучших инструментов машинного обучения. TensorFlow - это библиотека программного обеспечения с открытым исходным кодом для численного расчета с использованием графиков потока данных [1]. Существует прямая зависимость между скоростью обучения нейронной сети и точностью предсказания. Для ускорения процесса обучения используется вычислительная мощность видеокарты, а именно технология CUDA. CUDA- это архитектура параллельных вычислений от NVIDIA, позволяющая существенно увеличить вычислительную производительность благодаря использованию GPU (графических процессоров) [2].

Нейронная сеть — это громадный распределенный параллельный процессор, состоящий из элементарных единиц обработки информации, накапливающих экспериментальные знания и предоставляющих их для последующей обработки. Нейронная сеть сходна с мозгом с двух точек зрения:

– знания поступают в нейронную сеть из окружающей среды и используются в процессе обучения;

– для накопления знаний применяются связи между нейронами, называемые синаптическими весами [3].

Для обучения нейронной сети необходимо большое количество публикаций и метаданных, таких как дата создания, автор, ключевые слова и другие.

Перед обучением данные делятся на 2 части: для обучения и для тестирования.

Опишем процедуру обучения нейронной сети. На вход нейронной сети подается матрица векторов *MV*, каждый вектор *V* которой содержит информацию о конкретном атрибуте публикации (например, авторе). Для формирования отдельного вектора *V* перед обучением необходимо получить словарь *D* всех значений атрибута публикации. Словарь *D* сортируется по убыванию и отбрасываются последние *N* значений, чтобы нейронная сеть не обучалась на редко

встречающихся элементах, и тем самым не ухудшалась точность классификации. Указанная выше процедура выполняется для каждого атрибута. Для каждого автора публикации, производится поиск в словаре *D*, если данный автор найден, то под индексом найденного автора в вектор *V* ставится единица, иначе - ноль. Таким образом, заполняются все вектора матрицы *MV*.

Выходной вектор *R* описывает количество просмотров через заданный промежуток времени и состоит из единственного дробного числа, находящегося в диапазоне от нуля до единицы. Единица означает максимальное количество просмотров, в данном исследовании выбрано 50 миллионов.

Нейронная сеть состоит из 4 слоев (входной, два промежуточных и выходной слой). На промежуточных слоях используется функция активации *LeakyReLU*, на выходном слое применяется функция *softmax*. Между всеми слоями, кроме последнего, используется нормализация данных.

После обучения нейронной сети загружаются тестовые данные, и выполняется процедура тестирования. Далее на основании полученных векторов нейронная сеть предсказывает популярность статей через заданный промежуток времени. Данные обрабатываются и сохраняются в excel для анализа.

Обучение сети производилось на более чем 100 000 текстов, что занимает от 4 до 16 часов, в зависимости от глубины обучения и точности результата. В результате нейронная сеть способна предсказать количество просмотров с точностью в 75%. Верным считается ответ, находящийся в диапазоне +/-200 000 просмотров от ответа. Максимальное количество просмотров при обучении составляло 48 034 576. Коэффициент корреляции для массивов ответов и предсказанных значений составляет 0,3. Это означает, что между входными и выходными данными есть зависимость. Подобрав более точно входные данные или параметры нейронной сети, можно увеличить точность классификации.

Список литературы

[1]. Library for numerical computation using data flow graphs[Электронный ресурс]. – Режим доступа: https://www.tensorflow.org/. – Дата доступа: 15.03.2018г.

[2]. Параллельные вычисления CUDA[Электронный ресурс]. – Режим доступа: http://www.nvidia.ru/object/cuda-parallel-computing-ru.html. – Дата доступа: 15.03.2018г.

[3]. Хайкин, С. Нейронные сети: полный курс, 2-е издание/ С. Хайкин. — М. : Издательский дом «Вильямс», 2006. — 1104 с.

PREDICTIVE ANALYSIS SYSTEM FOR CLASSIFICATION OF TEXT DOCUMENT COLLECTIONS

A. KALOSHA
Master student, Department of Computer Science BSUIR

M. MEDUNETSKI,
Doctor of Technical Sciences
Professor

M. HORONEKO
Student BSUIR

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: andreikalosha@mail.ru

Abstract. We describe predictive analysis system for text document classification. Classification is handled by neural network by metrics set. We use number of views, likes and reposts as such metrics.

Key words: Big Data analytics, TensorFlow, CUDA, machine learning, neural networks.