УДК 004.896

СТАБИЛИЗАЦИЯ РАБОТЫ СЕТЕВОГО ПРИЛОЖЕНИЯ НА ОСНОВЕ ГЕНЕРИРУЕМЫХ ИМ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ, ИСПОЛЬЗУЯ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Козлов Е.А. (Университет ИТМО)
Научный руководитель – профессор, д.ф.-м.н. Попов И.Ю.
(Университет ИТМО)

В данной работе рассматривается и решается проблема стабилизации работы сетевого приложения на основе генерируемых им неструктурированных данных. Разработан и реализован соответствующий для этого алгоритм. Произведено тестирование решения, тем самым доказана его работоспособность.

Актуальность данной работы обусловлена активным развитием информационных технологий. Повсеместный процесс создания крупных информационных систем, отвечающих за разнообразные задачи, приводит к необходимости контроля и стабилизации их работоспособности. В настоящее время не существует универсального метода такого контроля систем, в связи с чем необходимо решить два глобальных вопроса – на основе каких данных, присущих любой информационной системе, будет проводиться анализ работы? Какие методы необходимо применить, чтобы сделать искомое решение максимально универсальным?

Каждая система в процессе своей работы генерирует так называемые лог-файлы. Это журналы, в которые в каждый момент времени записывается текущее состояние системы. Все такие файлы можно условно поделить на структурированные (например, в виде таблицы или JSON-структуры) и неструктурированные (текстовые данные, где каждая строчка имеет общий шаблон (уровень логирования), но вся информативная часть не имеет чётких размеров и содержания). Первый тип данных более простой для анализа и может как генерироваться данной системой, так и нет. Второй же тип данных более сложен для анализа и в целом присущ любой системе, в связи с чем возможность анализировать такие данные представляет наибольший интерес.

Однако не стоит забывать, что ввиду масштабов и сложности существующих информационных систем, размер генерируемых лог-файлов в единицу времени не позволяет производить их ручной мониторинг с целью контроля и стабилизации работы. В последние года в связи с активным развитием IT-индустрии как таковой, широко развивается и применяется математический аппарат, связанный с машинным обучением. Именно с помощью его методов построен универсальный алгоритм для решения поставленной задачи.

Главной идеей анализа неструктурированных данных является применение методов машинного обучения для обработки естественного языка. В первую очередь, это процесс конвертации данных из строчного формата в числовой. То есть построение процесса векторизации имеющихся данных. Однако искомое векторное пространство необходимо построить ещё и таким образом, чтобы слова, имеющие схожий контекст, обладали схожими координатами, и наоборот (таким образом, должна соблюдаться семантическая близость слов). После же того, как данные представляют из себя набор числовых координат, их необходимо объединить в группы по какому-то правилу, то есть кластеризовать (на языке машинного обучения – произвести обучение без учителя). Для этого за основу взяты два основных принципиально различных метода кластерного анализа – алгоритм K-средних (KMeans) и сканирования данных (DBSCAN). В первом случае изначально задаётся количество кластеров, в которое необходимо объединить данные; во втором – максимальное расстояние между соседними точками, исходя из которого алгоритм выдаёт получившееся число кластеров. Реализованы два этих алгоритма, произведён анализ полученных результатов и вывод о том, какой алгоритм лучше подходит для решения поставленной задачи.

Объединение слов в группы произведено, то есть финальный этап обработки данных завершён. Теперь необходимо уметь на основе таких данных делать выводы о текущей работе информационной системы. Для этого используются нейронные сети, а именно – рекуррентные. Краткая идея их архитектуры состоит в том, что таким сетям для расчёта следующего значения важен порядок, в котором данные поступали на вход. То есть появляется возможность обрабатывать серии событий во времени (последовательные цепочки). Однако простые рекуррентные сети (RNN) не могут долго “запоминать” информацию, в связи с чем существуют более продвинутые и сложные архитектуры нейронных сетей – сети с долгой краткосрочной памятью (LSTM) и управляемые рекуррентные блоки (GRU). В данной работе построены архитектуры данных видов рекуррентных сетей, получены результаты, произведён их анализ.

Результатом работы является механизм анализа, решающий поставленную задачу нахождения критических взаимосвязей в неструктурированных данных, предшествующих остановке работы системы.
Смоделированы синтетические данные, в которых была заведомо заложена ошибка работы сетевого приложения. Произведено тестирование разработанного решения на таких данных и получено заблаговременное выявление заложенной ошибки, то есть доказана работоспособность построенной модели.

Козлов Е.А. (автор)

Попов И.Ю. (научный руководитель)