

УДК 004.942

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКИХ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ ПО ЗАДАНЫМ ПОВЕДЕНЧЕСКИМ ХАРАКТЕРИСТИКАМ

Деева И.Ю. (Университет ИТМО), **Лысенко А.В.** (Университет ИТМО)
Научный руководитель – к. т. н. Калюжная А.В.
(Университет ИТМО)

Прогресс в области исследований и анализа социальных сетей зависит от наличия данных для моделирования и анализа социального поведения. Существуют ограничения конфиденциальности при сборе, обмене или распространении новых данных из социальных сетей (таких как данные Facebook, Twitter или ВКонтакте). Проблема конфиденциальности данных в социальных сетях накладывает жесткие ограничения на использование этих данных. Вот почему создание синтетических наборов данных является актуальной задачей.

Работа по созданию синтетических данных с заданными свойствами в основном касается создания изображений (изображений, лиц и т. д.) и звуков (например, музыки). В то же время проблема формирования синтетических профилей людей мало изучена. Однако такое исследование могло бы позволить как создание искусственных профилей для различных служб и алгоритмов, так и взаимодействие различных аспектов личности человека. Нет также работ по созданию синтетических транзакционных профилей или психологических профилей. Трудность заключается в создании кросс-модальных профилей, в которых агрегируются данные различной природы (например, банковские данные, публикации и фотографии в социальных сетях, демографические данные). Предложены подходы, основанные на автоэнкодерах для одновременной генерации текста и изображений или восстановления недостающих частей. Вопрос о том, как определить, насколько искусственные профили близки к реальным и актуальным, также остается неясным.

Задача генерации синтетических пользовательских данных может быть сформулирована в терминах байесовской статистики. В этой работе мы рассматриваем задачу генерации синтетических данных для профилей пользователей социальной сети (количество друзей, количество постов, количество видео на странице и т. д.) как задачу нескольких модальностей. Изучение совместного представления этих модальностей должно дать более глубокие и полезные представления. Мы используем мультимодальный вариационный автоэнкодер (MVAE), чтобы генерировать данные профиля пользователя на основе поведенческих и социально-демографических показателей.

Целью данной работы было создание генеративной модели цифровых личностей, в частности профилей социальных сетей. Для этого в работе использовался набор вариационных автоэнкодеров, скрытое пространство которых организовано таким образом, чтобы генерировать новые реалистичные объекты. Автоэнкодеры были обучены вместе с функцией потерь, которая отвечает за идентичность скрытых представлений различных моделей, что позволяет восстанавливать недостающие части профиля из доступных.

В работе рассматриваются варианты построения композиции вариационных автоэнкодеров для создания психологического профиля (BigFive) пользователей и их профиля в социальной сети, а также рассматриваются метрики оценки качества для сгенерированных профилей, которые наиболее подходят для этой задачи.