Тематическое моделирование сообществ социальной сети ВКонтакте с использованием ARTM-модели

Горшков Сергей Сергеевич

Студент

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

E-mail: serggorsar@yandex.ru

Научный руководитель *— Сухомлин Владимир Александрович*

В настоящее время социальные сети являются одним из самых больших источников данных. В научном сообществе и бизнесе решаются различные задачи, связанные с анализом социальных сетей — выявление интересов пользователей для рекомендаций, составление графов друзей для поиска потенциальных знакомых и прочие. Сейчас активно исследуется связь между цифровым следом пользователя и его психологическими характеристиками, личными достижениями, к примеру, связь между оценками обучающихся и характеристиками их профиля в социальной сети [1].

Одной из наиболее важных характеристик профиля в социальной сети являются подписки пользователя (группы, сообщества, в которых он состоит). Очевидно, что подписки коррелируют с интересами пользователя и потому вносят существенный вклад в его цифровой портрет. Подавляющее большинство описанных выше задач решается в настоящее время с помощью методов машинного обучения, и различные характеристики профиля используются в качестве признаков объектов. Учитывать в качестве признаков сообщества нетривиально ввиду их сложной структуры, использование же непосредственно уникальных номеров групп (подписан пользователь на каждую из групп или нет, в том числе с различными взвешиваниями оценок) не вносит никакого значимого вклада, что было проверено автором экспериментально.

Предлагаемое решение проблемы использования информации о группах — использование в качестве признаков распределения тем в сообществах пользователя. Для каждой подписки можно выделить одну или несколько тем, к которым относится контент группы, с помощью инструмента тематического моделирования для содержимого сообщества. В рамках исследовательского проекта рассматривались группы в социальной сети ВКонтакте, на которые подписаны более 10 студентов Томского государственного университета, представившие свои профили в ВКонтакте. Таковых сообществ оказалось 7042.

С помощью API сайта vk.com для каждой группы были выгружены название, описание и 50 последних записей (рекламные записи были отброшены). Далее эта информация была объединена в текстовое описание для группы, причем для решения проблемы отсутствия текстовой информации в постах в группе в первом приближении использовалась следующая эвристика — в результирующих наборах слов название группы и описание входили четыре раза, т.к. обычно тема описания и название группы коррелируют с содержанием. Далее, из текста были удалены знаки препинания, слова приведены в начальную форму в нижнем регистре, а также удалены стоп-слова.

Получившаяся коллекция документов была использована в качестве датасета для модели ARTM [2] с добавлением разреживающих регуляризаторов для матриц встречаемости терминов в теме и тем в документе. Использовалась только одна модальность – текстовая, в будущем планируется добавить в качестве модальностей хэштеги и изображения. Коэффициенты регуляризации выбирались перебором по сетке исходя из интерпретируемости тем, оценивалась перплексия, на основании значений которой выбиралось количество итераций работы алгоритма, и выделялись самые значимые слова в теме. В результате обучения модели коэффициент разреженности (доля нулей) в обеих матрицах составила примерно 0.9, количество тем составило 40. Удалось классифицировать 87% от исходного числа групп, отказы в классификации связаны, как правило, с очень маленьким числом слов в выборке или объясняются узкоспециализированностью тем. При этом распределение числа документов в образованных кластерах похоже на логнормальное, что ожидаемо. В лидерах по числу тем оказались группы про музыку, любовь, фильмы, афиша, историю.

Таким образом, каждой группе можно поставить в соответствие номер темы (или несколько тем с соответствующими весами) и в дальнейшем использовать распределение тем по всем подпискам пользователя как признаки, отражающие интересы пользователя.

Литература

1. Ихсанов И. Р., Шахова И. С. Применение методов машинного обучения для выявления взаимосвязи академической успеваемости и данных профиля социальной сети // Электронные библиотеки. 2019. Т. 22, № 2. С. 95–118.
2. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455, № 3. С. 268–271.