УДК: 004.942

Название:

Исследование интересов пользователей социальной сети на основе публичных постов и подписок

Авторы:

Утеуов А.К., Университет ИТМО, г. Санкт-Петербург;

Калюжная А.В., Университет ИТМО, г. Санкт-Петербург

Научный руководитель: Калюжная А.В, Университет ИТМО, г. Санкт-Петербург

Тезис доклада:

Целью работы является разработка подхода для исследования интересов пользователей социальных сетей, которые проявляются в различии публичной информации, которую пользователь публикует о себе и интересами пользователя, которые он напрямую не публикует.

В работе использовался анонимизированный корпус постов 10 тысяч пользователей российской социальной сети, количественная информация о профиле (число друзей, подписок, возраст) и информация об их подписках на сообщества (название сообществ, количество подписчиков). Суммарное количество уникальных слов после токенизации для датасетов users_posts и users_subscr 1.3 млн слов и 920 тыс. соответственно.

На основе постов пользователей было проведено тематическое моделирование и выделены топ слова для каждой темы. По подпискам были получены вероятностные распределения по схожим темам. Текстовые данные были предобработаны: исключены небуквенные символы, удалены стоп-слова и конвертированы в формат представления Vowpal Wabbit. Для каждого пользователя имелось два набора текстов: объединенные посты на стене профиля и объединенные в один текст названия групп, на которые подписан пользователь. Для моделирования использовались реализации моделей PLSA и ARTM библиотеки BigARTM. Для большей разреженности матрицы «документы-темы» использовался разреживающий регуляризатор с параметром -10, количество тем 64, количество итераций 30. Вычисления были выполнены на высокопроизводительном блейд сервере.

Для каждого пользователя были получены два распределения по темам: публичное и приватное (по подпискам). Полученные темы были размечены, пример тем пользователя: путешествия, фильмы, музыка, игры, религия, культура, дети, фильмы, бизнес, спорт. Пример тем сформированным по подпискам пользователей: бизнес, спорт, дети, любители животных, кулинария, медицина. Затем была составлена матрица попарной схожести тем по топ 60 словам, отсортированные по вероятности принадлежности к теме, и рассчитана метрика Жаккара (мера совпадения множеств, далее “мера схожести”). Затем для каждого пользователя были добавлены значения этой метрики в зависимости от доминирующей приватной и публичной темы. Таким образом была получено значение различия интересов для каждого пользователя через меру схожести между публичной и приватной темой.

Было проведено сравнение публичной и приватной тем, по результатам которого только у 10% пользователей наблюдается их совпадение. Распределение количества пользователей по темам показывает, что существуют темы типичные для подписок пользователей, связанные с частными интересами и хобби. И также существуют общие темы, которые формируют темы в публичных постах пользователей.

Результатом работы является разработанный подход для проведения тематического моделирования интересов пользователей социальной сети, реализованные скрипты и проведенные эксперименты для указанных текстовых наборов. Данная методика позволит лучше предсказывать интересы пользователя в рекомендательных системах.