Решение проблемы холодного старта в рекомендательных системах с помощью анализа социальных сетей

А.В. Аникин

Самарский университет, Самара, Россия

Обоснование. Объем информации в интернете оценивается в настоящее время как 100 зеттабайт и продолжает расти [1]. Для ориентации пользователей в таких объемах информации создаются специальные рекомендательные системы. В настоящее время рекомендательные системы применяются на самых разнообразных программах, на сайтах и веб-ресурсах. Для их корректной работы нужна различная информация о предпочтениях пользователя. Чаще всего она собирается во время использования сайта — учитываются переходы пользователя на страницы, оценки, покупки, движения курсора и другие параметры.

Часто имеет место ситуация, называемая проблемой «холодного старта», когда еще не накоплено достаточное количество данных для корректной работы рекомендательной системы. Самые известные решения данной проблемы — это использование сглаженного среднего и доверительного интервала. Они имеют свои особенности применения и не дают существенного прироста в точности рекомендаций.

Цель — повысить качество рекомендаций на начальных этапах работы (уменьшить среднеквадратичную ошибку — RMSE) рекомендательной системы с помощью использования данных из социальных сетей пользователя.

Методы. В качестве социальной сети решено было использовать социальную сеть «ВКонтакте» как наиболее популярную и имеющую удобный API [2]. С помощью него можно получать большое количество информации о пользователе, но для работы системы понадобятся только информация со страницы (возраст, количество фото, уровень образования, интересы, жизненная позиция), отметки «нравится» на постах, содержащих ключевые слова (названия фильмов, жанров), подписки на группы.

Создаваемая рекомендательная система является гибридной [3, 4]. Для ее построения используются данные о фильме и о пользователе. Для фильма это поля: год, жанр, оценка пользователя и коэффициент заинтересованности пользователя, который высчитывается на основе данных из социальной сети. Данный коэффициент увеличивается в случае, если у пользователя есть связанные с фильмом отметки «нравится», подписки на группы и собственные посты. Также о самом пользователе сохраняются данные о возрасте, образовании и любимых жанрах.

Затем система была протестирована в приложении по подбору фильмов, которое выступило основой. На основе тестовых данных получились разнообразные результаты для разных пользователей, прежде всего связаны они были с количеством информации о пользователе в социальной сети «ВКонтакте». Среднеквадратичная ошибка вычислялась по формуле:

$$RMSE = \sqrt{\frac{\sum_{i \in n} (R_i - \overline{R}_l)^2}{n}},$$

где $RMSE$ — средняя квадратичная ошибка; n — количество значений (оценок); R_i — предсказываемое значение; $\overline{R}_l$ — реальное значение (оценка пользователя).

Результаты. Пользователь, имеющий заполненную информацию о себе, несколько сотен подписок, 20 из которых оказали влияние на результат, и около тысячи отметок «понравилось», 230 из них имели конечное влияние, показал такие результаты среднеквадратичной ошибки:

1) 1,82 — *RMSE* без использования данных из социальных сетей;
2) 1,64 — *RMSE* с использованием данных из социальных сетей.

В другом случае пользователь с незаполненным профилем, имеющий 20 подписок, 3 из них имели конечное влияние, и 18 отметок «понравилось», которые почти не оказали влияние на результат.

В связи с таким маленьким набором данных уменьшение среднеквадратичной ошибки оказалось несущественным:

1) 1,82 — *RMSE* без использования данных из социальных сетей;
2) 1,81 — *RMSE* с использованием данных из социальных сетей.

Выводы. Описанный в работе способ помогает добиться прироста в точности рекомендаций на тестовых данных. Уменьшение среднеквадратичной ошибки зависит от количества данных в социальной сети, связанных с объектом рекомендаций. Планируется апробировать гипотезу и на реальных данных и пользователях, запустив приложение в общем доступе.

Ключевые слова: рекомендательные системы; проблема холодного старта; анализ социальных сетей.

Список литературы

1. Vopson M.M. The information catastrophe // AIP Advances. 2020. Vol. 10, N 8. ID 085014. doi: 10.1063/5.0019941
2. dev.vk.com [Электронный ресурс]. API ВКонтакте [дата обращения: 24.05.24]. Режим доступа: https://dev.vk.com/ru/api/overview
3. Оболенский Д.М., Шевченко В.И. Обзор современных методов построения рекомендательных систем — на основе контента и гибридные системы. В кн.: Сборник статей Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых: «Мир компьютерных технологий»; 05–09 апреля 2021; Севастополь. Севастополь: ФГАОУВО Севастопольский государственный университет, 2021. С. 151–156.
4. Phuong N.D., Thang L.Q., Phuong T.M. A graph-based method for combining collaborative and content-based filtering. В кн.: PRICAI 2008: Trends in artificial intelligence. PRICAI 2008. Lecture notes in computer science. Vol. 5351 / Ho T.B., Zhou Z.H., editors. Springer, Berlin, Heidelberg. 2008. P. 859–869. doi: 10.1007/978-3-540-89197-0_80

Сведения об авторе:

Арсений Васильевич Аникин — студент, группа 6304-010302D, Институт информатики и кибернетики; Самарский университет, Самара, Россия. E-mail: ars.anikin.2003@gmail.com

Сведения о научном руководителе:

Александр Владимирович Благов — кандидат технических наук, доцент; Самарский университет, Самара, Россия. E-mail: blagov@ssau.ru