ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ МЕТОДОВ ПОСТРОЕНИЯ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Асташкевич М.Г.

Волорова Н. А.- заведующая кафедрой информатики

Развитие информационных технологий, в частности, увеличение вычислительной мощности техники и удешевление хранения данных, позволяет собирать и анализировать огромные объемы информации о пользователях, в том числе данные о предпочтениях и поведении пользователей. Эти данные представляют большую ценность для различных организаций. Например, поисковые гиганты, такие как Google, Yandex и другие, используют такие данные о том или ином пользователе сети для ранжирования результатов поисковых запросов, а также для более эффективного подбора показываемых рекламных объявлений. Также в качестве примера можно привести эксперимент исследователей из бизнес-школы Уорика (Warwick Business School), которые предположили возможность прогнозирования поведения рынка ценных бумаг на основе анализа открытых данных сервиса Google Trends. В результате исследователи смогли увеличить портфель акций на 8% за две недели - что является достаточно впечатляющим результатом, учитывая, что исходные данные брались из открытых источников. Таким образом, преследуемая автором цель заключается в исследовании современных подходов к прогнозированию и построению рекомендаций на основе данных, находящихся в открытом доступе.

Задача об анализе данных о поведении и предпочтениях пользователя с целью последующей обработки и построения определенных прогнозов относительно дальнейшего его поведения не нова. Крупные интернет-сервисы десятилетиями предлагают пользователям такую функциональность - Last.fm предлагает музыку, Netflix предлагает фильмы, Amazon предлагает товары и т.д. Способность предложить пользователю то, что ему нужно, может сильно повлиять на успех коммерческого проекта.

Входные данные таких алгоритмов можно обобщить к следующему виду: пользователь x совершает действие r_xy по отношению к объекту y. Действие в данной трактовке - это просмотр информации об объекте, “лайк”, выставленная оценка, оставленный комментарий - словом, все, что показывает интерес пользователя к данному объекту. Таким образом образуется входная матрица рейтингов - большая и сильно разреженная.

Рекомендательные системы, которые и занимаются обработкой такого рода информации, делятся, в общем, на два вида - основанные на содержимом рекомендательные системы и методы коллаборативной фильтрации, причем коллаборативная фильтрация являет собой более универсальный подход. Простейшие алгоритмы основываются на группировке пользователей по оцененным объектам, или на группировке объектов по оценившим их пользователям. В обоих случаях сценарий схожий: определить способ группировки пользователей, рекомендовать пользователю объекты, интересные его группе пользователей (и наоборот - для рекомендательных систем, основанных на группировке по объектам). Среди коллаборативных рекомендательных систем выделяется алгоритм SVD, который работает несколько другим образом и признан на данный момент самым перспективным и наиболее точным.

SVD (Singular Value Decomposition), переводится как сингулярное разложение матрицы. В теореме о сингулярном разложении утверждается, что у любой матрицы $\mathbf{A}$ размера $n \times m$ существует разложение в произведение трех матриц: U, Σ и V^T. Зная о существовании такого разложения, можно использовать принципы машинного обучения для того, чтобы подобрать такие матрицы U и V, которые получат наименьшую погрешность при прогнозировании рейтинга для известных элементов матрицы r (обучающая выборка) и прогнозировать с их помощью события в будущем. Для определения погрешности можно использовать, например, среднеквадратичное отклонение (RMSE). Для того, чтобы алгоритм не только находил наилучшее решение для известных данных, но также имел хорошие результаты предсказывая будущие события, нужно к данной метрике погрешности добавить регуляризатор. Регуляризация заключается в том, что оптимизируется не просто ошибка, а ошибка плюс некоторая функция от параметров (например, норма вектора параметров). Это позволяет ограничить размер параметров в решении, уменьшает степень свободы модели. Таким образом задача о построении рекомендательной системы сводится к задаче о нахождении таких матриц U и V, для которых метрика ошибки плюс регуляризатор минимальны. С этим можно справится, например, с помощью алгоритма градиентного спуска.

Таким образом, в проведенном исследовании были рассмотрены основные современные алгоритмы построения рекомендаций, разработан алгоритм тестирования рассмотренных алгоритмов на основе анализа информации о комментариях пользователей крупного белорусского новостного сайта. Были также рассмотрены существующие программные решения для решения подобных задач. Также были определены возможные дальнейшие направления исследования в данной области.

Список использованной литературы:

1. Алексей Кострикин. Введение в алгебру. Часть 1. Основы алгебры / Алексей Кострикин. — 1-е изд. — Физико-математическая литература, 2001. — 272 с.
2. Владимир Зорич. Математический анализ / Владимир Зорич. — 1-е изд. — МЦНМО, 2007. — 1458 с.
3. Константин Воронцов. Машинное обучение (курс лекций) / Константин Воронцов.

Библиотека БГУИР