Ли Хань
Студент 2 курс магистратуры,
факультет «Информатика, искусственный
интеллект и системы управления»
Московский государственный технический
университет им. Н.Э. Баумана

БАЗИРУЮЩИЙСЯ НА ПРОДАЖАХ ТОВАРОВ И МНОГИХ ДЕЙСТВИЯХ ПОЛЬЗОВАТЕЛЕЙ АЛГОРИТМ СОВМЕСТНОЙ ФИЛЬТРАЦИИ

Аннотация. Объясняется потенциальное влияние уровней продаж товаров на пользователей, выбирается формула вычисления сходства Пирсона для дальнейших расчетов. Через множество действий пользователей с товаром предоставляются различные веса оценок для пользователей. Результаты экспериментов показывают, что в сравнении с традиционными алгоритмами совместной фильтрации данный алгоритм повышает точность вычислений сходства.

Abstract. The potential impact of product sales levels on users is explained, and a Pearson similarity calculation formula is selected for further calculations. Through multiple user actions with the product, different rating weights are provided to users. The experimental results show that this algorithm improves the accuracy of similarity calculations compared to traditional collaborative filtering algorithms.

Ключевые слова: Рекомендуемые системы, Совместная фильтрация, Действия пользователей, Параметры самых продаваемых товаров.

Keywords: Recommendation system, collaborative filtering, user behavior, best-selling product parameters

1.Введение

Алгоритм рекомендаций на основе товаров сначала вычисляет степень сходства между товарами, определяя несколько соседних групп, связанных с целевым товаром, затем, рассчитывая оценки разных товаров, получает предсказанные значения оценок для неоцененных товаров. Подтверждено, что основанные на товарах алгоритмы рекомендаций проявляют лучшие показатели, помогая решать проблемы разреженности и масштабируемости, с которыми сталкиваются традиционные алгоритмы [1]. Однако, с критической точки зрения, точность рекомендаций на основе товаров зависит от объема данных и вычислительных ресурсов; обычно рекомендации на основе пользователей оказываются более точными [2].

Алгоритмы совместной фильтрации являются одним из самых популярных методов рекомендаций. Их концепция заключается в следующем: если разные пользователи ставят похожие оценки для одного и того же товара, то их интересы к другим товарам также должны быть близки [3].Однако в эпоху больших данных объем данных, необходимых для вычисления рекомендаций, значительно увеличивается, что усложняет алгоритмы. Открытость интернета и проблемы безопасности усиливают настороженность пользователей по отношению к интернет-информации, что приводит к проблемам с разреженностью данных в принципах рекомендаций и снижает точность алгоритмов. Поэтому многие исследователи предложили различные улучшенные алгоритмы [3].Например, с помощью категориальных меток в одном исследовании [4] использовались байесовские сети для точного вычисления условных вероятностей, что улучшало актуальность алгоритма. В другом исследовании [5], рассматривая предпочтения пользователей через оценку и атрибуты товаров, был предложен алгоритм UPPPCF, который компенсирует недостатки традиционных алгоритмов, рассчитывающих сходство только по оценкам. Однако упомянутые работы также имеют свои недостатки: во-первых, они не учитывают потенциальное влияние уровней продаж товаров на

выбор пользователя. Кроме того, вычисление сходства только между парами пользователей не всегда точно отражает их схожесть; многие пользователи оценивают товары не только по своим предпочтениям, но также учитывают отзывы окружающих или похожих пользователей. То есть, уровни продаж товаров могут потенциально влиять на суждения и выборы пользователей.

В данной работе учитывается влияние уровней продаж товаров на пользователей, улучшается алгоритм на основе множественных действий пользователей, основываясь на различных весах оценок для формирования рекомендательных списков.

2.Традиционные алгоритмы совместной фильтрации

Основная идея алгоритма состоит в том, что сначала ищется группа соседей, схожих с целевым пользователем, а затем выбираются проекты, которые могут понравиться целевому пользователю, на основе интересов похожих пользователей. Методы вычисления сходства пользователей включают:

①Косинусное сходство

Считается, что оценки пользователей на товары представляют собой векторы в n-мерном пространстве. Косинус угла между векторами пользователей можно использовать для измерения степени сходства между ними. Чем больше результат косинусного сходства, тем более схожими являются два пользователя. Косинусное сходство вычисляется, как показано в формуле (1).

$$sim(u,v) = cos\left(\overrightarrow{u},\overrightarrow{v}\right) = \frac{\sum_{i=1}^{n} r_{ui} * r_{vi}}{\sqrt{\sum_{i=1}^{n} r_{ui}^2} \times \sqrt{\sum_{i=1}^{n} r_{vi}^2}}. \quad (1)$$

Где u и v соответственно представляют собой векторы двух пользователей в n-мерном пространстве.

②Модифицированное косинусное подобие

Традиционный метод расчета косинусного сходства не учитывает, что разные пользователи имеют разные стандарты оценивания товаров; некоторые оценивают более сдержано, а другие – более широко. Модифицированное косинусное сходство учитывает этот фактор, вводя среднюю оценку пользователей и вычитая ее для улучшения результата. Модифицированное косинусное сходство вычисляется, как показано в формуле (2).

$$sim(u,v) = \frac{\sum_{i\in I_{u,v}}(r_{ui} - \overline{r_u}) \times (r_{vi} - \overline{r_v})}{\sqrt{\sum_{i\in I_u}(r_{ui} - \overline{r_u})^2} \times \sqrt{\sum_{i\in I_v}(r_{vi} - \overline{r_v})^2}}. \quad (2)$$

Где I_u – это множество товаров, оцененных пользователем u, I_v – множество товаров, оцененных пользователем v, $I_{u,v}$ – множество товаров, оцененных совместно пользователями u и v, $\overline{r_u}$ – средняя оценка пользователя u для товаров из этого множества, $\overline{r_v}$ – средняя оценка пользователя v для товаров из этого множества.

③Коэффициент корреляции Пирсона (Pearson)

Метод коэффициента корреляции Пирсона основан на расчете линейной зависимости между оценочными векторами пользователей. Он определяется через отношение ковариации между двумя переменными оценок пользователей и стандартных отклонений. В общем случае коэффициент корреляции Пирсона варьируется в диапазоне -1, 1; если пользователи ставят схожие оценки одному и тому же товару, то они более схожи.Коэффициент корреляции Пирсона вычисляется, как показано в формуле (3).

$$sim(u,v) = \frac{\sum_{i\in I_{u,v}}(r_{ui} - \overline{r_u}) \times (r_{vi} - \overline{r_v})}{\sqrt{\sum_{i\in I_{u,v}}(r_{ui} - \overline{r_u})^2} \times \sqrt{\sum_{i\in I_{u,v}}(r_{vi} - \overline{r_v})^2}}. \quad (3)$$

Где $I_{u,v}$ – это множество товаров, оцененных совместно пользователями u и v, $\overline{r_u}$ – средняя оценка пользователя u для товаров из этого множества, $\overline{r_v}$ – средняя оценка пользователя v для товаров из этого множества.

3.Описание алгоритма

3.1Представление множественных характеристик поведения пользователя

Характеристики пользователя могут включать в себя различные данные и в настоящее время представлены предпочтениями пользователя. Характеристики предпочтений отражают потребности пользователя в определенных товарах или типах товаров. Электронная коммерция хранит много данных о поведении пользователей, таких как записи, оставленные при просмотре товаров, записи о коллекциях любимых товарах, записи о покупках, информация об оценке товара после покупки и т.д. Эти данные отражают субъективное отношение пользователя к определенному типу товаров и его приоритеты при их выборе, поэтому они могут использоваться для описания предпочтений пользователя. Наиболее точным ответом на симпатии и антипатии пользователя к товарам являются рейтинги. Таким образом, характеристики предпочтения пользователя к определенному продукту могут быть выражены его рейтингом. Например, если пользователь отбирает какой-то предмет и выражает интерес к нему, интерес может быть отражен в балле. Если пользователь покупает предмет после его выбора, он балл становится более высоким. На основе анализа записи покупок пользователя, создается модель персонализированных рекомендаций, базирующаяся на рейтингах. Количество просмотров определенного предмета также может отражать интерес пользователя, и этот интерес также выражается в баллах. Чем большее количество раз элемент просматривается, тем выше должен быть рейтинг продукта после просмотра всей коллекции товаров. Для системы рекомендаций, основанных на совместной фильтрации, базирующейся на рейтингах и записях просмотра, данные о характеристиках пользователя очень важны. Данные рейтинга являются самым основным индикатором характеристик пользователя и лучшим выражением пользовательских предпочтений. По сравнению с другими материалами, данных о пользовательском рейтинге не так много, что может привести к проблемам с разреженностью данных при построении матриц.

Основной деятельностью веб-сайтов электронной коммерции является предоставление потребителям услуг онлайн-покупок, и пользователи веб-сайтов электронной коммерции имеют определенные персонализированные предпочтения. В данной работе товарам, с которыми взаимодействовал пользователь, присваивается соответствующий вес, исходя из уровня интереса, проявленного пользователем к странице данного товара, добавления его в корзины покупок, покупки и комментированию сделанной покупки. На основании действий пользователя формируется рейтинг пользователя для товаров, который использует веса, значения которых присваиваются по правилам, указанным в таблице 1.

Таблица 1

Представление весовых показателей данных характеристик пользователя.

Данные о поведении пользователя	Вес
Избранное	1
В корзину	2
Комментарии	2
Заказ на покупку	3

3.2Уровень продаж товаров

На платформах электронной коммерции Самые продаваемые товары часто рекомендуются каждому пользователю, однако они не могут адекватно отразить интересы и предпочтения пользователей.

В общем случае, согласно правилу 80/20: если 80% пользователей оценили товар, независимо от оценок, это означает, что уровень продаж товара очень высок, и нет смысла рекомендовать этот товар пользователю. Даже если пользователь ничего не оценивал, он, скорее всего, уже знаком с этим товаром. Если только 20% пользователей оценили товар, то этот товар соответствует принципу длинного хвоста, и соответствующие рекомендации принесут пользователю лучшее качество опыта [6].

С другой стороны, если товар очень популярен, вероятность его рекомендации пользователю может быть в n раз выше, чем у непопулярного товара. Таким образом, чтобы более точно оценить степень сходства пользователей, в данном разделе предлагается новый метод, включающий параметр уровня продаж в расчет сходства пользователей, чтобы смягчить негативное влияние популярных товаров на результаты рекомендаций. Параметр модификации уровня продаж определяется, как указано в формуле (4).

$$\frac{1}{\lg(1+|N(i)|)}. \tag{4}$$

Внедряя формулу 3.1 в коэффициент Пирсона, мы получаем модифицированную формулу расчета сходства, как показано в формуле (5).

$$sim(u,v)=\frac{\sum_{i\in I_{u,v}}(r_{ui}-\overline{r_u})\times(r_{vi}-\overline{r_v})\times\frac{1}{\lg(1+|\mathrm{N}(i)|)}}{\sqrt{\sum_{i\in I_{u,v}}(r_{ui}-\overline{r_u})^2}\times\sqrt{\sum_{i\in I_{u,v}}(r_{vi}-\overline{r_v})^2}}. \tag{5}$$

4.Результаты экспериментов и анализ

4.1Экспериментальная среда

Язык программирования – Python, экспериментальная платформа – Google Colab, данные для экспериментов взяты из набора данных Retailrocket recommender system на сайте kaggle. Этот набор данных был собран с реального интернет-магазина. Данные являются сырыми, то есть без предварительной обработки. Данные по действиям, такие как клики, добавление в корзину, покупки и т. д., представляют собой данные о взаимодействии пользователей, собранные за 4,5 месяца. Пользователи могут создавать три типа событий: "просмотр", "добавление в корзину" или "покупка". Всего зарегистрировано 2,756,101 события, включая 1,407,580 просмотров, 69,332 добавлений в корзину и 22,457 покупок, созданных 11,719 пользователями после очистки данных.

4.2Критерии оценки

Существует множество методов оценки производительности алгоритмов совместной фильтрации, например, средняя абсолютная ошибка (MAE), среднеквадратичная ошибка (RMSE), полнота (Recall), точность (Precision), F1Score и т. д. В данной работе основными критериями оценки являются полнота (Recall), точность (Precision) и F1Score.

4.3Результаты экспериментов и их анализ

Результаты экспериментов представлены на Рисунках 1 и 2.

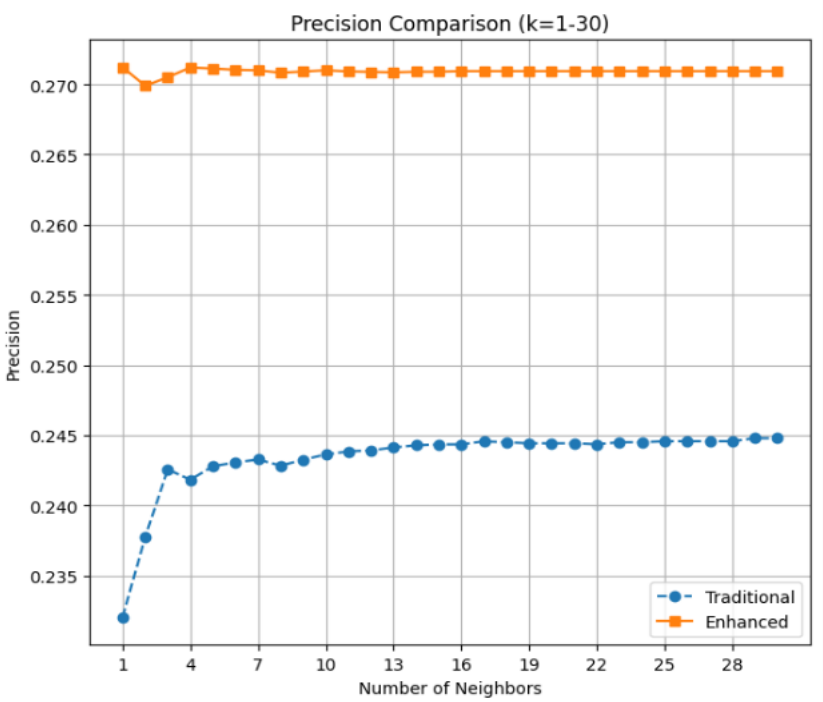

Рисунок 1 Comparison of Precision

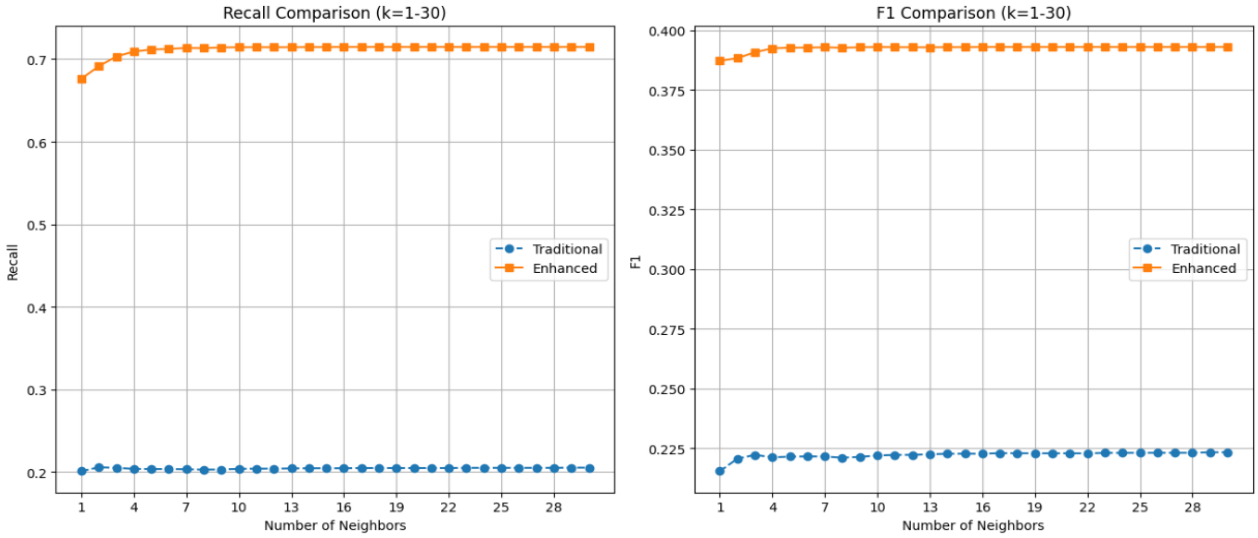

Рисунок 2 Comparison of *Recall* and *F1*

На Рисунке 1 видно, что по мере увеличения числа соседей модифицированный алгоритм совместной фильтрации сначала уменьшал, а затем увеличивал точность, стабилизировавшись позже, в то время как традиционный алгоритм совместной фильтрации сначала увеличивал, а затем снижал точность, также достигая устойчивого уровня, но модификация обеспечивает более высокие значения точности по сравнению с традиционным алгоритмом. Кроме того, как видно на Рисунке 2, модифицированный алгоритм показывает более высокие значения полноты и F1, что свидетельствует о том, что модифицированный алгоритм эффективно улучшает результаты рекомендаций традиционных алгоритмов.

5.Пути дальнейшего развития исследований

Хотя предложенный в статье модифицированный алгоритм в определенной степени повысил точность рекомендаций и может хорошо решать проблемы низкой точности рекомендаций, разреженности данных и масштабируемости, остаются некоторые недостатки и области для улучшения. В дальнейшем работа будет продолжена по нескольким направлениям:

В данной статье при расчете сходства пользователей введен штрафной фактор для популярных товаров, что помогает пользователю целевой группы более точно находить соседей, тем самым повышая производительность рекомендаций. Однако в реальных условиях можно наблюдать, что интересы пользователей не остаются стабильными, поэтому для исследования закономерностей их изменений и взаимосвязи с временными факторами исследование в будущих моделях должно учитывать изменения интересов пользователей со временем.

Предложенный алгоритм испытан только на одном наборе данных; в будущих исследованиях могут быть включены другие наборы данных, такие как MovieLens, Epinions и т.д., для проверки эффективности данного алгоритма.

Список литературы:

1. Хань.К, Ван.Ю, Шэнь.Ч. Исследование автоматического извлечения тем китайского текста на трех уровнях / Журнал китайской информатики – 2001. 20-26.

2. Ма.И, Ван.Ю, Сюй.Г. Метод извлечения тем китайского текста, основанный на частоте совместного появления иероглифов / Исследования и разработки в области компьютерных наук – 2003. 874-878.

3. Вэнь.Ю, Вэнь.Х, Сюй.Ж. Извлечение знаний на основе инновационных точек / Журнал информатики – 2005. 664-668.

4. Чжан.Ю, Гун.Л, Ван.Ю. Автоматическое извлечение тематических предложений текста на основе интегрированного метода / Журнал Шанхайского транспортного университета – 2006. 771-774.

5. Хэ.В, Ван.Ю. Извлечение тематических предложений веб-текста на основе графа отношений предложений / Современные библиотечные и информационные технологии – 2009. 57-61.

6. Хао.Л, Ван.Ц. Алгоритм рекомендаций TopN на основе популярности товаров и совместной фильтрации / Компьютерная инженерия и проектирование – 2013. 3497-3501.

7. Набор данных рекомендательной системы Retailrocket – 2022. – Метод доступа: https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset

8. Хань.Л. Алгоритмы доверительного рекомендования и фильтрации поиска друзей в социальных сетях / Университет Яншань – 2012.