Шныра Богдан Олегович
Студент II-го курса магистратуры
кафедра автоматизированные системы управления
ФГБОУ ВО «Донецкий национальный технический университет»
e-mail: mi_shnyra@mail.ru
г. Донецк, Донецкая Народная Республика, Россия

Андриевская Наталия Климовна
кандидат технических наук
кафедра автоматизированных систем управления
ФГБОУ ВО «Донецкий национальный технический университет»
г. Донецк, Донецкая Народная Республика, Россия

ИСПОЛЬЗОВАНИЕ КОНТЕНТНОЙ ФИЛЬТРАЦИИ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ДЛЯ ВЕБ-ПЛАТФОРМЫ

УДК 004.8

Аннотация:
Шныра Б.О., Андриевская Н.К. Использование контентной фильтрации рекомендательной системы для веб-платформы. *В современном мире развитие интернет-технологий и веб-платформ привело к необходимости создания эффективных инструментов обработки и анализа больших объемов графических данных. На основе анализа визуального контента, представленного пользователями, формируются персонализированные предложения для пользователей. Особое внимание занимает контентная фильтрация. Для ее реализации произведено извлечение признаков текста посредством токенизации и преобразования в эмбеддинги, а также извлечение визуальных признаков с использованием предобученной модели CNN ResNet-50. В рамках контентной фильтрации были описаны процессы формирования положительных и отрицательных пар для обучения модели. Разработана архитектура нейросетевой модели, определены метрики оценки качества модели, а также предложены рекомендации по ее дальнейшему улучшению.*

Ключевые слова: *рекомендательные системы**,** сверточные нейросети, обработка текста, обработка изображения, прогнозирование.*

Постановка проблемы. Рекомендательные системы (РС) являются важной областью исследований в сфере информационных технологий, направленной на предоставление пользователям персонализированных предложений товаров, услуг или контента, соответствующих их индивидуальным предпочтениям и потребностям. С ростом объёмов доступной информации и ассортиментного разнообразия в электронной коммерции, социальных сетях и цифровых медиа, задача эффективного фильтрации и представления релевантной информации становится всё более актуальной [1].

Основная проблема, с которой сталкиваются рекомендательные системы, связана с необходимостью преодоления информационной перегрузки и предоставления точных и своевременных рекомендаций.

Основные методы решения проблемы. Один из методов решения проблемы информационной перегрузки в рекомендательных системах — это контентная фильтрация.

Основная задача метода контентной фильтрации заключается в нахождении похожих элементов на основе анализа их содержимого. Контентная фильтрация (Content-Based Filtering) базируется на характеристиках самих элементов.

Например, если пользователь часто смотрит фильмы определенного жанра или с определенными актерами, система будет рекомендовать ему похожие фильмы. Этот метод требует анализа и обработки данных о самих элементах, таких как жанры, актеры, режиссеры и т.д. Контентная фильтрация предполагает, что пользователи будут заинтересованы в элементах, схожих с теми, которые они уже оценили положительно. Она анализирует свойства объектов, такие как текстовые описания, метаданные и ключевые слова, чтобы выявить схожие элементы. Это позволяет системе рекомендовать пользователю материалы, аналогичные тем, которые он уже оценил положительно [2].

Разработка моделей для контентной фильтрации. Первым этапом была выполнена предварительная обработка данных – это процесс анализа данных, который включает преобразование необработанной информации в удобный для анализа формат и дальнейшего использования [3]. После предварительной обработки данных выполняется векторизация текстов, так как алгоритмы машинного обучения предназначены для работы с числовыми данными, и необходимо выполнить преобразование текста в числовой вектор признаков [4].

Для извлечения текстовых признаков из заголовков и описаний была использована предобученная модель BERT [5]:

$$Tokens = Tokens(text), \quad (1)$$

где text — это исходный текст, Tokens — это список токенов, представленных в виде числовых идентификаторов.

После процесса токенизации токены поступали в модель BERT, которая преобразовала их в эмбеддинги:

$$Embedding_{CLS} = BERT(Tokens)[0], \quad (2)$$

где BERT(Tokens) — это выход модели BERT для заданных токенов;

а[0] — это выбор эмбеддинга[CLS], который находится на первой позиции в выходном векторе.

После получения эмбеддингов, заголовки и их описания были объединены для формирования единого вектора признаков:

$$d_c = d_h + d_d, \quad (3)$$

где $d_n -$ вектор заголовка, $d_d -$ вектор описание.

Объединение выполняется конкатенацией:

$$Embedding_{combined} = \left[Embedding_{title}; Embending_{descriptions}\right], \quad (4)$$

где $Embedding_{title}$ и $Embending_{descriptions} -$ эмбеддинги заголовка и описания;

[;] обозначает операцию конкатенации векторов.

Для извлечения визуальных признаков использована предобученная модель ResNet-50 [6]. Для создания единого представления элемента, текстовые и визуальные признаки объединяются:

$$d_{combined} = d_{text} + d_{image}, \quad (5)$$

где $d_{text} -$ текстовые признаки, $d_{image} -$ визуальные признаки.

Объединение признаков текста и изображения происходит с помощью операции конкатенации:

$$combined\_features = torch.cat\big((combined\_text, image\_features), dim = 1\big), \quad (6)$$

где dim=1 указывает на то, что конкатенация выполняется по размерности признаков [7].

Для обучения модели на основе сходства были созданы пары элементов. Положительные пары сформированы из элементов, которые идентичны, то есть представляют собой один и тот же объект. Эти пары помечаются меткой 1. В данной работе выбран подход формирования пар на основе uid пользователя:

$$(x_i, x_j) \text{ где } UID(x_i) = UID(x_j) \rightarrow label = 1, \quad (7)$$

где UID(x)— уникальный идентификатор элемента x.

Отрицательные пары сформированы из элементов, которые различны, и помечаются меткой 0. Формирование отрицательных пар сделана случайным:

$$(x_i, x_j) \text{ где } UID(x_i) \neq UID(x_j) \rightarrow label = 0, \quad (8)$$

где UID(x)— рандомный уникальный идентификатор элемента x.

Процент положительных и отрицательных пар контролируется параметром positive_ratio:

$$N_{positive} = N_{negative} = \frac{N}{2}, \quad (9)$$

где N – это общее количество пар.

Недостатком формирование случайным образом отрицательных пар является то, что многие из них будут слишком легко различимы моделью, что не способствует эффективному обучению.

Для предсказания схожести между парой признаков применена нейронная сеть, которая состоит из нескольких слоев, обеспечивающих обработку и анализ входных данных. Архитектура модели описывается следующими компонентами:

Входной слой принимает объединенные признаки двух элементов, что приводит к следующему количеству измерений:

$$input\_dim = 3584 \rightarrow total\_input\_dim = 2 * input = 7168, \quad (10)$$

где input_dim = **3584** - это количество признаков, извлеченных из одного элемента;

total_input_dim = 2 × input_dim = 7168 - пара элементов, общее количество входных измерений увеличивается в два раза.

Модель включает в себя несколько скрытых слоев, которые выполняют функции обработки и обучения.

Первый полносвязный слой содержал 512 нейронов и для него функция активации ReLU (Rectified Linear Unit) определяется как:

$$ReLU(x) = \max(0, x) \quad (11)$$

Dropout слой применялся с вероятностью p=0.3 для предотвращения переобучения, что означает, что 30% нейронов будут случайным образом отключены во время тренировочного этапа.

Второй полносвязный слой имел 128 нейронов и такую же функцию активации ReLU, как и первый. Второй Dropout слой также применялся с вероятностью p=0.3.

Выходной слой состоял из одного нейрона, который использовал функцию активации Sigmoid для предсказания вероятности схожести между элементами. Функция Sigmoid определяется как:

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (12)$$

Таким образом, выходной слой предсказывал вероятность схожести p в диапазоне от 0 до 1:

$$p = \sigma(z) \text{ где } z = W * h + b, \quad (13)$$

где W — это вес выходного слоя, h — выход предыдущего слоя, b — смещение.

Для задач бинарной классификации [8], где необходимо было предсказать, является ли пара элементов схожими (1) или не схожими (0), применялась функция потерь бинарной кросс-энтропии. Она измеряет расхождение между истинными метками y и предсказанными вероятностями p. Формально, функция потерь BCE определяется следующим образом:

$$L(y, \rho) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\rho_i) + (1 - y_i)\log(1 - \rho_i)), \quad (14)$$

где N— общее количество образцов в батче, yi — истинная метка (0 или 1) для i-го образца, p_i — предсказанная вероятность схожести для i-го образца.

Результаты тестирования построенных моделей. В ходе обучения модели было проведено 5 эпох. На рисунке 1 представлена динамика изменения значения функции потерь (Loss) во время обучения модели. Наблюдается устойчивое снижение значения Loss с 0.6244 на первой эпохе до 0.3934 на пятой эпохе.

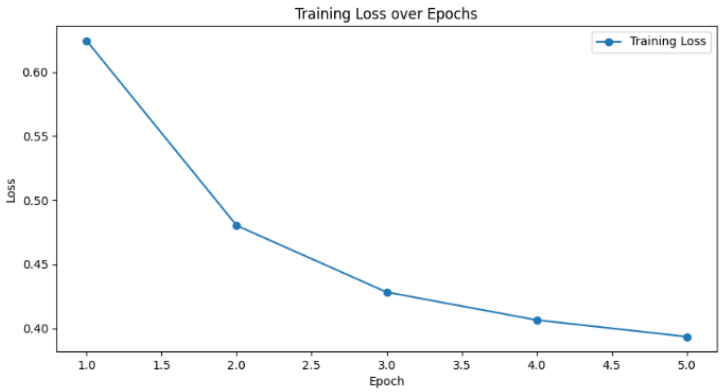

Рисунок 1 - График функции потерь на тренировке

На рисунке 2 отображены изменения метрик Precision (точность), Recall (полнота) и F1-Score модели на валидационной выборке в течение пяти эпох обучения [9].

Динамика метрик указывает на первоначальное повышение точности модели с последующим балансированием точности и полноты. Увеличение Recall и F1-Score свидетельствует о том, что модель становится способной выявлять все большее количество релевантных примеров, сохраняя при этом приемлемый уровень точности.

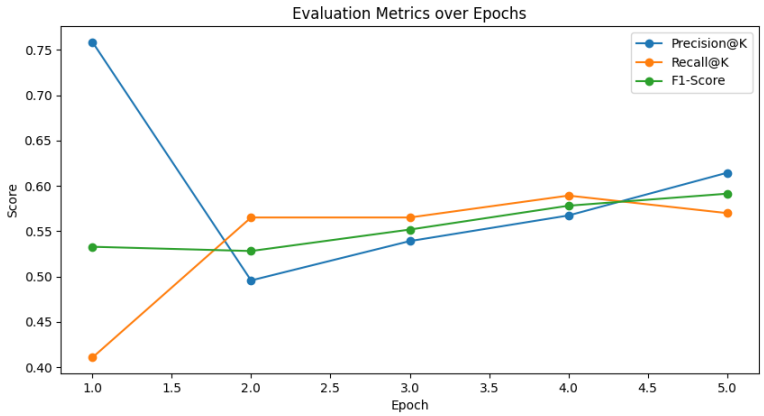

Рисунок 2 – Графики оценочных метрик при валидации

Анализ результатов моделирования.

Показатели метрик качества для пяти эпох составили от 55% до 60%, что является средним результатом.

В дальнейших исследованиях, в первую очередь, следует увеличить количество эпох для достижения более приемлемого результата.

Для улучшения показателей качества построенной модели при генерации отрицательных пар для обучения следует использовать Hard Negative Mining вместо случайного формирования негативных пар.

Этот метод сосредоточен на отборе отрицательных пар, которые модель может легко классифицировать как положительные, что позволит улучшить её способность различать схожие, но разные элементы. Для этого будут использоваться такие метрики, такие как косинусное сходство и евклидово расстояние [10].

Необходимо выбрать негативные пары с высокими значениями сходства. Эти пары будут считаться «жесткими негативами» и представляют собой примеры, которые модель может легко перепутать с положительными [11]. Отрицательная пара (x_i,x_j) считаются жестким негативом, если:

$$f(x_i, x_j) \geq threshold \text{ и } UID(x_i) \neq UID(x_j), \quad (15)$$

где f(xi,xj)— это функция сходства, а threshold— заранее установленный порог.

Заключение. Полученные результаты демонстрируют эффективное обучение модели, подтверждённое стабильным снижением функции потерь на тренировочных данных. Несмотря на первоначальное снижение метрики Precision, последующее восстановление и общий рост метрик Recall и F1-Score свидетельствуют о достижении баланса между точностью и полнотой предсказаний. Но для более приемлемого результата в дальнейшем будет увеличено эпох, так же для улучшения показателей качества построенной модели при генерации отрицательных пар для обучения будет использоваться метод Hard Negative Mining.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Введение в рекомендательные системы справочник [Электронный ресурс] \ Режим доступа: https://www.researchgate.net/publication/227268858_Recommender_Systems_Handbook. Дата обращения: 30.10.2024.
2. Рекомендательные методы системной фильтрации [Электронный ресурс] \ Режим доступа: https://underskyai.ru/post/rekomendatelnyie_metodyi_sistemnoy_filtratsii_kontentnaya_filtratsiya_ee_plyusyi_i_minusyi. Дата обращения: 30.10.2024.
3. Вовченко, В. О. Формирование датасета для решения задач машинного обучения / В.О. Вовченко, В.А. Светличная, Н.К. Андриевская // Информатика и кибернетика. – 2023. – № 2(32). – С. 5-12.
4. Андриевская, Н.К. Разработка алгоритмов предобработки информации для прогнозных моделей ИСППР управления закупками / Н.К. Андриевская, Т.В. Мартыненко // Информатика и кибернетика. – 2023. – № 3(33). – С. 11-18.
5. BERT: предварительное обучение [Электронный ресурс] \ Режим доступа: https://arxiv.org/abs/1810.04805 Дата обращения: 03.11.2024.
6. Шныра, Б.О. Использование сверточных нейросетей для анализа визуального контента веб-платформы / Б.О. Шныра, Н.К. Андриевская // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2024) : сб. материалов XV Междунар. науч.-техн. конф. в рамках X Междунар. Науч. форума ДНР; Т.2 / Ред. кол.: Аноприенко А. Я. (пред.); Васяева Т. А.; Карабчевский В. В. [и др.]; от. ред. Р. В. Мальчева. – Донецк : ДонНТУ, 2024. – С. 690-695.

7. Объединение признаков Torch [Электронный ресурс] \ Режим доступа: https://www.geeksforgeeks.org/how-to-join-tensors-in-pytorch/ Дата обращения: 03.11.2024.
8. Решение задач бинарной классификации [Электронный ресурс] \ Режим доступа: https://habr.com/ru/companies/infopulse/articles/307150/ Дата обращения: 07.11.2024.
9. Оценка модели Accuracy, Precision, F1 [Электронный ресурс] \ Режим доступа:https://pythonru.com/baza-znanij/metriki-accuracy-precision-recall Дата обращения: 07.11.2024.
10. Евклидово расстояния и косинус схождение [Электронный ресурс] \ Режим доступа:https://www.baeldung.com/cs/euclidean-distance-vs-cosine-similarity Дата обращения: 07.11.2024.
11. Жестко отрицательный майнинг [Электронный ресурс] \ Режим доступа: https://cvexplained.wordpress.com/2020/07/15/2-9-hard-negative-mining/ Дата обращения: 07.11.2024.

Shnyra Bogdan Olegovich
Student of the II-rd course of the undergraduate
Department of Automated Control Systems
Donetsk National Technical University
e-mail: mi_shnyra@mail.ru
Donetsk, Donetsk People's Republic, Russia

Andrievskaya Natalia Klimovna
Candidate of Engineering Sciences
Department of Automatic Control Systems
Donetsk National Technical University
Donetsk, Donetsk People's Republic, Russia

USING CONTENT FILTERING RECOMMENDATION SYSTEM FOR WEB PLATFORM

Annotation:

Shnyra B.O., Andrievskaya N.K. Using content filtering of a recommender system for a web platform*. In the modern world, the development of Internet technologies and web platforms has led to the need to create effective tools for processing and analyzing large volumes of graphic data. Based on the analysis of visual content submitted by users, personalized offers for users are formed. Particular attention is paid to content filtering. For its implementation, text features were extracted by tokenization and transformation into embeddings, as well as visual features were extracted using a pre-trained CNN ResNet-50 model. Within the framework of content filtering, the processes of forming positive and negative pairs for training the model were described. The architecture of the model was also developed, the metrics used to assess its quality were determined, and recommendations for its further improvement were proposed.*

Keywords: *recommender systems, convolutional neural networks, text processing, image processing, forecasting.*