DOI 10.37882/2223-2966.2024.06.09

СИСТЕМАТИЗАЦИЯ ДАННЫХ ДЛЯ РАЗРАБОТКИ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ПОВЕДЕНЧЕСКИХ ПАТТЕРНОВ АУДИТОРИИ ИНТЕРНЕТ-РЕСУРСОВ

SYSTEMATIZATION OF DATA FOR THE DEVELOPMENT OF A MODEL FOR PREDICTING BEHAVIORAL PATTERNS OF THE AUDIENCE OF INTERNET RESOURCES

D. Buhonov

Summary. Modern Internet resources strive to optimize user experience to attract and retain audiences. Therefore, visitor behavioral data is a valuable source of information. Their analysis allows us to understand the preferences and needs of users, which can then be used as a basis for various research and development of software solutions.
The work proposes the use of a mathematical model based on the results of an analysis of behavioral factors of users of a large media resource to optimize its operation.

Keywords: behavioral data, statistical data, Information technology, CTR, headlines, website optimization.

Бухонов Дмитрий Олегович
Аспирант, ФГБОУ ВО «Ульяновский государственный технический университет»;
Ведущий разработчик, ООО «ТТК ДИДЖИТАЛ»
d.buhonov@yandex.ru

Аннотация. Современные интернет-ресурсы стремятся оптимизировать пользовательский опыт для привлечения и удержания аудитории. Поэтому поведенческие данные посетителей являются ценным источником информации. Их анализ позволяет понять предпочтения и потребности пользователей, что затем можно использовать как базу для различных исследований и разработки программных-решений.
Работа предлагает использование математической модели, основанной на результатах анализа поведенческих факторов пользователей большого медиаресурса, для оптимизации его работы.

Ключевые слова: поведенческие данные, статистические данные, информационные технологии, CTR, заголовки, оптимизация сайтов.

Использование статистических данных выполняет важную функцию в процессе оптимизации работы ресурсов и повышения эффективности стратегий маркетинга. Анализ этих данных предоставляет ценную информацию о поведении пользователей, позволяя выявить её предпочтения и потребности, а также определить сильные стороны сайта. Это дает возможность улучшить пользовательский опыт, подстроить контент под интересы аудитории и оптимизировать рекламные кампании с учётом реальных предпочтений посетителей [3].

Целью данной работы является работа над информацией, которая станет основой модели прогнозирования, использующей статистические данные для улучшения работы сайтов. Для её реализации необходимо:

- Разобрать основные возможности использования статистических данных об активности пользователей со статейными заголовками в рамках улучшения работы интернет-ресурса;
- Определить влияние поведенческих факторов на процентную метрику показов и прочтений;
- Проанализировать возможности оптимизации сайтов и привлечения целевой аудитории с помощью программных решений, основанных на этих данных.

В качестве основного метода исследования был выбран метод наименьших квадратов, поскольку он является одним из наиболее оптимальных инструментов, позволяющих увеличить коэффициент достоверности предсказания.

Эффективность интернет-ресурсов становится важным инструментом в современном образовании, исследованиях и повседневной жизни. Различные статьи и исследования обсуждают роль и возможности, которые предоставляют интернет-ресурсы в различных областях: от обучения иностранным языкам до научных исследований.

Интернет-ресурсы активно внедряют инструменты для персонализации контента и улучшения пользовательского опыта [3]. Интеграция цифровых инструментов используется для оптимизации рекламных алгоритмов в ритейле, а также для исследований, трендов, влияющих на будущее цифровой экономики и бизнеса [1].

Основываясь на актуальных направлениях работы веб-ресурсов, можно выделить несколько путей применения пользовательских данных:

1. **Сегментация аудитории**. Статистические данные позволяют выделить ключевые сегменты аудито-

рии, что помогает создать персонализированный контент и улучшить взаимодействие с пользователями [2].

2. **Оптимизация контента.** Анализ статистики помогает понять, какой контент на сайте наиболее популярен и востребован, что позволяет создавать более привлекательный и целенаправленный материал [1].
3. **Улучшение пользовательского опыта.** Статистические данные о поведении пользователей на сайте помогают выявить слабые места интерфейса и оптимизировать его для удобства посетителей [1].
4. **Оценка эффективности рекламы.** Анализ данных о трафике и поведении после рекламной кампании позволяет оценить её результативность и корректировать стратегию [2].
5. **Прогнозирование поведения клиентов.** Статистические модели на основе пользовательских данных помогают предвидеть предпочтения пользователей и их дальнейшие действия, что требуется для улучшения персонализированных маркетинговых стратегий [2], повышения посещаемости и равномерного распределения трафика в рамках одного сайта.

Для реализации математической модели в данной работе использовались основные показатели поведенческих данных на выбранном ресурсе: показы, клики, загрузки заголовка.

Обоснование метода обучения модели и выбранных данных

В качестве показателя эффективности материала (статьи на сайте) выбрано соотношение просмотров к кликам по заголовкам (далее: win_rate).

Математическая модель представлена линейной регрессией. Линейная регрессия представляет собой статистический метод, используемый для моделирования отношений между зависимой переменной (в данном случае, win_rate) и одной или несколькими независимыми переменными (признаками: avg_views, avg_clicks, даты обновления, создания и другие). Для обучения модели используется метод наименьших квадратов для обучения модели линейной регрессии (LeastSquares).

Для определения логики взаимосвязей между классами, методами и техническими классами, а также для выбора данных и обучения модели использовалась диаграмма (рис. 1).

Для решения задачи прогнозирования win_rate на основе поведенческих данных пользователей сайта был выбран подход, использующий линейную регрессию. Данный выбор обусловлен несколькими факторами, включая структуру данных, их доступность и понятность результатов.

С датасета осуществляется выборка средних значений собранных показателей действий пользователей, на основе чего выполняется обучение модели. Обобщается информация по следующим показателям:

— просмотры;
— уникальные просмотры;
— клики;
— уникальные клики;
— максимальное и минимальное количество загрузок;
— даты просмотров.

Для построения модели был использован ArrayDataset, который позволяет работать с массивами данных, что является удобным для предварительной обработки информации перед обучением модели. Данный подход класс также предоставляет возможность более гибко манипулировать данными и признаками, а также упрощает подготовку данных для обучения.

LeastSquares был выбран в качестве метода линейной регрессии для обучения модели. Этот метод позволяет

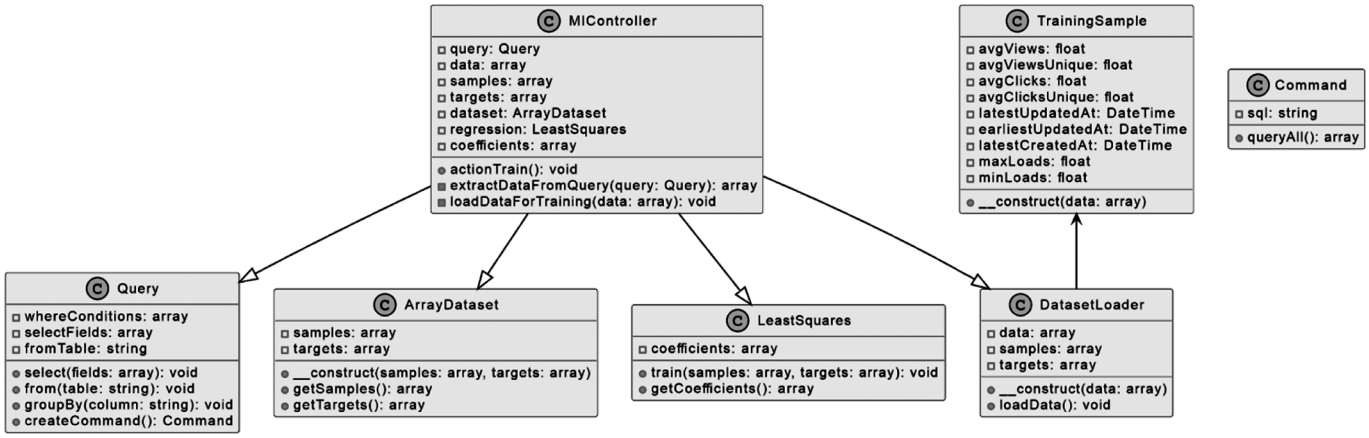

Рис. 1. Визуализация диаграммы классов

аппроксимировать зависимость между независимыми переменными (признаками) и зависимой переменной (win_rate) линейной функцией. При этом используется метод наименьших квадратов для минимизации разницы между фактическими и предсказанными значениями.

Выбранные данные представляют собой агрегированную информацию о статьях на сайте, включая средние значения просмотров, кликов, даты обновления и создания, а также максимальные и минимальные значения загрузок. Эти данные были выбраны на основе их релевантности и предполагаемого влияния на пользовательскую активность. Например, количество просмотров и кликов, а также даты обновления и создания, могут дать представление о популярности и актуальности статьи, влияя таким образом на win_rate.

Эти коэффициенты регрессии представляют значимость каждого признака в прогнозировании win_rate. Давайте проанализируем результаты:

0 (avg_views): -0.0595
1 (avg_views_unique): 0.0697
2 (avg_clicks): 0.9123
3 (avg_clicks_unique): -1.0383
4 (latest_updated_at): 0.2347
5 (earliest_updated_at): 1.0398
6 (latest_created_at): -5.0225
7 (max_loads): 8.8691
8 (min_loads): 0.0023

Анализ результатов

- avg_views и avg_views_unique, хоть и имеют небольшие коэффициенты, всё же оказывают некоторое влияние на win_rate. Это может указывать на то, что среднее количество просмотров, хоть и не является ключевым фактором, все же оказывает некоторое влияние на вовлеченность пользователей.
- avg_clicks и avg_clicks_unique имеют значительное влияние на win_rate, что подчеркивает важность активности пользователей. Большое количество кликов положительно влияет на win_rate, в то время как уникальность кликов имеет негативное воздействие.
- latest_updated_at и earliest_updated_at также оказывают сильное положительное воздействие на win_rate, подчеркивая важность свежести и актуальности контента для пользователей.
- latest_created_at имеет отрицательное влияние, что может означать, что слишком свежий контент может ещё не успеть набрать необходимую активность.

Остальные признаки, близкие к нулю, могут означать, что они имеют менее существенное влияние на win_rate.

Исходя из анализа коэффициентов регрессии, можно сделать вывод о факторах, оказывающих сильное влияние на win_rate статьи. Количество кликов, их уникальность, актуальность статьи по датам обновления и создания играют ключевую роль в определении пользовательской активности. Эти результаты могут помочь оптимизировать сайт и контент для увеличения win_rate.

Полученные результаты позволяют увидеть важность активности пользователей, актуальности контента и, в некоторой степени, его привлекательности (просмотров) для определения win_rate и дальнейшего анализа его в рамках работы с интернет-ресурсом.

ЛИТЕРАТУРА

1. Або-Рашед Кнаан. Использование инструментов веб-аналитики для улучшения посещаемости сайта // Научный результат. Информационные технологии. — Т.5. — №2, 2020
2. Кедров С.А., Кузнецов С.О. Исследование групп пользователей Интернет-ресурсами методами анализа формальных понятий и разработки данных (Data Mining) // Бизнес-информатика. — ФГАОУ ВО «НИУ «ВШЭ», Москва. — 2007. — С. 45–57.
3. Нурутдинов Т.А. Архитектура и программное обеспечения для сбора пользовательских событий (clickstream) и их последующего анализа // Universum: технические науки: электрон. научн. журн. — 2023. — 12(117). URL: https://7universum.com/ru/tech/archive/item/16541 (дата обращения: 19.12.2023).

Журнал «Современная наука: актуальные проблемы теории и практики»