МАТЕМАТИЧНІ МОДЕЛІ ДЛЯ ПРОГНОЗУВАННЯ ВІДВІДУВАНОСТІ ВЕБСАЙТІВ

Іванов Дмитро
доктор технічних наук
Грас Андрій
здобувач вищої освіти магістерського рівня
Усата Олена
кандидат педагогічних наук
Кафедра комп'ютерних наук та інформаційних технологій
Житомирський державний університет імені Івана Франка

Інтернет-трафік відіграє ключову роль у функціонуванні багатьох онлайн-ресурсів. Для управління вебсайтами та прийняття рішень, зокрема в маркетингу та бізнесі, важливо мати інструменти для прогнозування майбутніх змін у кількості відвідувачів. Математичні моделі дозволяють здійснювати такі

прогнози, спираючись на наявні дані. Ця стаття аналізує основні підходи до прогнозування вебтрафіку. Одним із найпростіших підходів до прогнозування відвідуваності є лінійна регресія. Вона дозволяє знаходити залежність між змінною трафіку та факторами, які можуть на нього впливати, такими як час, рекламні кампанії, зміни в контенті сайту. Лінійна регресія описується рівнянням: $y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\varepsilon$, де y – залежна змінна (в даному випадку, кількість відвідувачів), $x_1,x_2,\ldots, x_n$ – незалежні змінні (час, обсяг реклами тощо), $\beta_0,\beta_1,\ldots,\beta_n$ – коефіцієнти моделі, ε – похибка. Цей підхід є розширенням лінійної регресії та дозволяє моделювати нелінійні залежності. Вона ефективна при складних взаємозв'язках між факторами і трафіком. Часові ряди – це послідовність спостережень, виконаних у часі. Для вебсайтів часові ряди дозволяють прогнозувати трафік на основі минулих даних. ARIMA є популярним методом для аналізу часових рядів і добре підходить для прогнозування відвідуваності, коли дані мають сезонні або циклічні тенденції. ARIMA комбінує три компоненти: авторегресію (AR), інтегровану частину (I) і компонент ковзного середнього (MA).

Основне рівняння ARIMA: $y_t=\alpha+\beta_1y_t-1+\cdots+\beta_py_t-p+\theta_1\varepsilon_t-1+\cdots+\theta_q\varepsilon_t-q+\varepsilon_t$,y_t, де y_t – прогнозована величина, α — константа, β_i— коефіцієнти автокореляції, θ_i – коефіцієнти ковзного середнього, ε_t – похибка.

Деякі вебсайти мають чітко виражену сезонність трафіку, наприклад, зростання кількості відвідувань під час свят або розпродажів. У таких випадках корисно використовувати моделі, які враховують сезонні коливання. Машинне навчання пропонує більш гнучкі методи для прогнозування трафіку, які можуть враховувати складні нелінійні залежності та взаємодії між факторами. Дерева рішень дозволяють будувати моделі, що ділять дані на підгрупи на основі певних характеристик, що дозволяє отримати прогноз на основі історичних даних.

Штучні нейронні мережі є потужним інструментом для прогнозування трафіку, особливо в умовах великої кількості даних. Вони здатні навчатися складним закономірностям і тенденціям. Одним з найпопулярніших варіантів для таких завдань є рекурентні нейронні мережі (RNN), які зберігають інформацію про попередні стани системи.

Для ефективного прогнозування відвідуваності вебсайтів важливо правильно вибрати модель залежно від характеристик даних. Наприклад, для короткострокових прогнозів можуть бути ефективні часові ряди, тоді як для довгострокових прогнозів доцільніше застосовувати нейронні мережі. На практиці часто використовують комбіновані підходи, такі як моделі гібридного прогнозування, що поєднують кілька різних методів.

Моделювання вебтрафіку має низку обмежень. Серед них: висока волатильність даних, вплив зовнішніх факторів, таких як зміни в алгоритмах пошукових систем або раптові зміни у поведінці користувачів. Крім того, прогнози можуть бути неточними у разі недоступності достатньої кількості історичних даних.

Ідентифікація математичної моделі відвідуваності веб-сайту здійснюється в два етапи:

1. Ідентифікація параметрів динаміки системи: на першому етапі на основі спостережень щодо оновлення контенту (функція якості) ідентифікуються параметри динамічної системи. Це включає використання функцій зміни активності відвідувачів сайту.

2. Ідентифікація параметрів загальної відвідуваності: на другому етапі ідентифікуються параметри моделі загальної відвідуваності веб-сайту, що дозволяє описати загальну активність користувачів.

Процес ідентифікації здійснюється на основі критерію мінімізації середньоквадратичної похибки між фактичною активністю користувачів та прогнозованою моделлю. Для цього використовується модифікований градієнтний метод Левенберга-Марквардта, що забезпечує ефективний пошук локального мінімуму. Проте, цей метод потребує добре підібраних початкових наближень для уникнення проблем з локальними мінімумами.

Оскільки кожен прояв активності відвідувачів є унікальним через людський фактор, модель відвідуваності повинна бути адаптивною. Це означає, що параметри моделі мають уточнюватися в реальному часі, що дозволяє реагувати на змінні умови контенту та поведінки користувачів.

Для побудови точних моделей, зокрема для системи, необхідно отримати хоча б чотири точки для ідентифікації. Це включає початкові точки на етапі зростання відвідуваності та кілька точок, прогнозованих на основі апріорних оцінок.

Функції якості оновлення контенту мають складну нелінійну природу, що ускладнює їх прогнозування. Однак аналіз різних реалізацій цих функцій показує наявність певних шаблонів – інтервали зростання та спадання. Це дозволяє класифікувати процеси зростання відвідуваності на короткотермінові та довготермінові.

Дослідження показують, що друга похідна функції якості контенту може бути наближена лінійною моделлю, що спрощує прогнозування параметрів у періоді зростання. На основі цієї властивості розроблені рекурентні співвідношення для покрокового прогнозування значень функції якості в контрольних точках.

Процес прогнозування включає ініціалізацію початкових значень першої та другої похідних функції якості, а також багатокроковий прогноз, який враховує зміни цих похідних. Прогнозування триває до тих пір, поки похибка не перевищує допустимого порогу.

Для початкової ідентифікації параметрів системи використовуються сіткові методи, що дозволяють отримати наближення коефіцієнтів моделі. Наприклад, коефіцієнт a4a_4a4, який відповідає за балансування, визначається на вузлах рівномірної сітки W4W_4W4, що покриває певний діапазон значень.

Прогнозування відвідуваності вебсайтів є складним, але важливим завданням для багатьох сфер діяльності. Існує безліч математичних моделей, які можуть бути ефективними у різних контекстах. Вибір відповідної моделі залежить від специфіки даних та цілей прогнозування. Машинне навчання та моделі часових рядів є одними з найперспективніших підходів у цій сфері.

Список використаних джерел

1. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control. Wiley. DOI: 10.1111/jtsa.12194.
2. Пасічник Н.Р. Математичні моделі відвідуваності вебсайтів та методи їх ідентифікації : дис... канд. техн. наук : 01.05.03 / Тернопільський національний економічний університет. Тернопіль, 2014. 178 с.
3. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 2018. URL: https://otexts.com/fpp2/
4. Zhang G. P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 2003. P. 159-175.