УДК 519.7:378.147

МАТЕМАТИЧНА МОДЕЛЬ ДИНАМІКИ ВІДВІДУВАНОСТІ ТЕМАТИЧНИХ ВЕБ-САЙТІВ ТА МЕТОДИ ЇЇ ІДЕНТИФІКАЦІЇ

Н.Р. Пасічник, Р.М. Пасічник, М.П. Дивак

Тернопільський національний економічний університет,

natalia.pasichnyk@gmail.com, roman.pasichnyk@gmail.com, mdy@tneu.edu.ua

У статті запропоновано метод параметричної ідентифікації моделі процесу короткотермінового росту відвідуваності Веб-сайтів та метод параметричної ідентифікації моделі процесу довготермінового росту відвідуваності Веб-сайтів. В роботі розглядається неавтономна модель відвідуваності Веб-сайтів, яка включає модель динаміки фактора впливу на основі модифікованої системи звичайних диференціальних рівнянь Моно

Ключові слова: Веб-сайт, параметрична ідентифікація, неавтономна модель, система диференціальних рівнянь Моно.

The article offers a method for parametric identification of the model of the short-term growth period of the Website's attendance and a method for parametric identification of the model of the long-term growth period of the Website's attendance. The work considers non-autonomous model of the Website's attendance, which contains the model of impact factor's dynamics basing on the modified system of differential equations by Mono.

Keywords: Web-site, parametric identification, non-autonomous model, system of differential equations by Mono.

В статье предложен метод параметрической идентифікации модели процеса краткострочного роста посещаемости Веб-сайтов и метод параметрической идентификации модели процеса долгострочного роста посещаемости Веб-сайтов. В работе рассматривается неавтономная модель посещаемости Веб-сайтов, которая содержит модель динамики фактора влияния на основании модифицированной системы обычных дифференциальных уравнений Моно.

Ключевые слова: Веб-сайт, параметрическая идентификациія, неавтономная модель, система дифференциальных уравнений Моно.

Вступ

Заходи по підвищенню відвідуваності Веб-сайту реалізуються протягом тривалого часу згідно відпрацьованих методик. Як в реалізації методики, так і в реакції на неї аудиторії Веб-сайту, значну роль відіграє суб'єктивний фактор. Оцінити результативність цієї діяльності зачасту можна лише після завершення її активної стадії. Маючи прогноз динаміки процесу на початковій його стадії, було б доцільно мати можливість скоригувати тактику реалізації даного виду діяльності або вчасно спланувати нову посилюючу активність. Це породжує необхідність побудови прогнозної моделі відвідуваності Веб-сайту.

Прогнозування відвідуваності Веб-сайтів вже частково досліджено в літературі, зокрема в роботах [1], [2]. Згадані роботи аналізують питання навігації користувачів по сторінках Веб-сайту з метою виявлення найчастіше

відвідуваних маршрутів а також питання вибору ефективних Веб-сайтів та Веб-сторінок для розміщення реклами. Більшість із зазначених підходів не дає рекомендацій про шляхи підвищення відвідуваності Веб-сайтів із низькою та середньою відвідуваністю. Хоча в дослідженнях А.М.Пелещишина [3] відвідуваність Веб-сайту пропонується підіймати в рамках Веб-холдингу за рахунок використання сайтів-донорів, однак значна частина користувачів позбавлена такої можливості.

Проведений аналіз літературири і практичних розробок дозволив побудувати загальну схему підтримки функціонування Веб-сайтів, представлену на рисунку1. Із Веб-сайтом взаємодіє його аудиторія А, формуючи його відвідуваність Y. Розробкою, підтримкою та модифікацією Веб-сайту займаються розробники *D* та служба підтримки *P* на основі суб'єктивних уявлень. Зазначений суб'єктивізм в розвитку Веб-сайту може завадити йому реалізувати потенційний рівень відвідуваності. Для цього необхідно підкріпити суб'єктивну інтуїцію працівників об'єктивними рекомендаціями, сформованими на основі відповідних інформаційних та математичних моделей.

Побудова напрямків поповнення контенту та розробки нових структурних елементів Веб-сайту може бути значно спрощена із використанням Веб-онтологій, побудова яких описана в попередніх дослідженнях [4],[5]. Крім того модель відвідуваностей Y дає можливість прогнозувати результати відвідуваності варіантів розвитку Веб-сайту ще до їх реалізації. Таким чином розробка методів ідентифікації математичних моделей для підтримки процесів структурування Веб-сайтів шляхом дослідження обсягів відвідуваності становить актуальну задачу, що розв'язується в даній роботі.

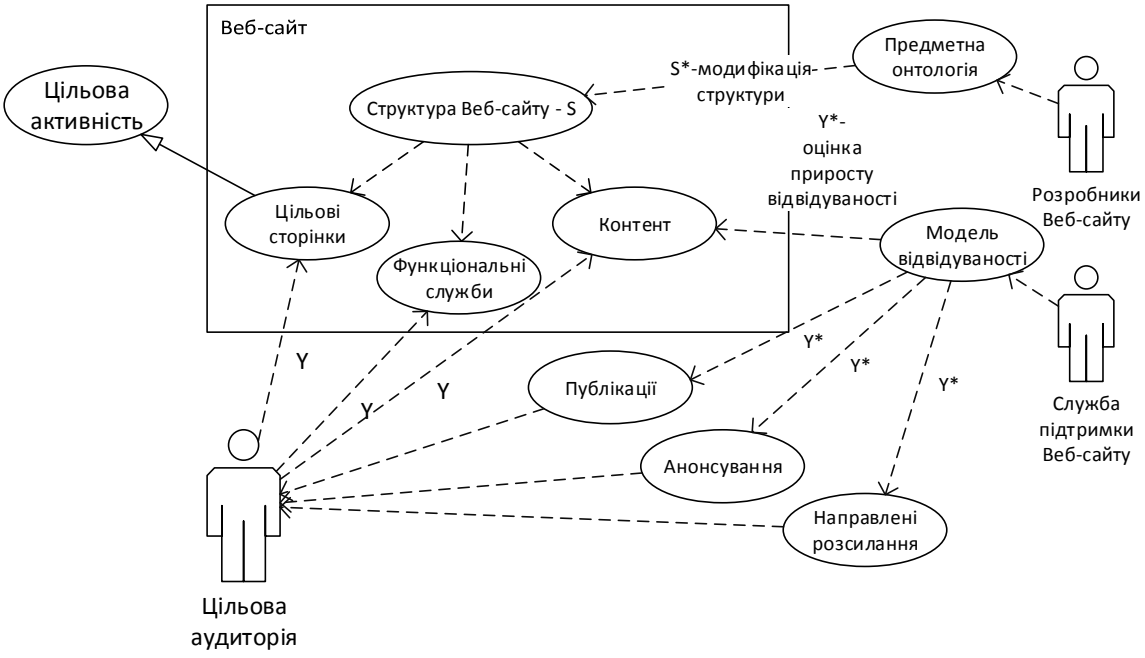

Рис1. Схема підтримки функціонування Веб-сайтів

Порівняльний аналіз відвідуваностей Веб-сайтів дозволив виявити узгодженість загальних характеристик відвідуваностей із відвідуваністями окремих тематик даного Веб-сайту (субсайтів), що на даний момент активно розвивалися або відповідали ключовим інтересам його аудиторії. В деяких випадках цю відвідуваність можна прогнозувати на основі оцінки стану бізнес-процесів об'єкта для якого створювався даний Веб-сайт.

1.Модель росту відвідуваності Веб-сайтів

Щоденна відвідуваність характеризується великою кількістю випадкових факторів, що робить проблематичним достатньо точне її прогнозування. В той же час на характеристики сайту впливають не окремі екстремальні, а усереднені показники. Тому в нашій моделі аналізуватимемо середньотижневі згладжені загальні відвідуваності Y та відвідуваності субсайтів X. При цьому використаємо кратне згладжування методом ковзаючого середнього із мінімальним шаблоном, яке на відміну від інших методів суттєво нівелює аномальні викиди, а не лише частково згладжує їх.

Для формального виділення інтервалів активності росту відвідуваності використовуємо критерій перевищення похідної відвідуваності деякого мінімального значення

$$Y'(t) > D_{min} \quad . \tag{1}$$

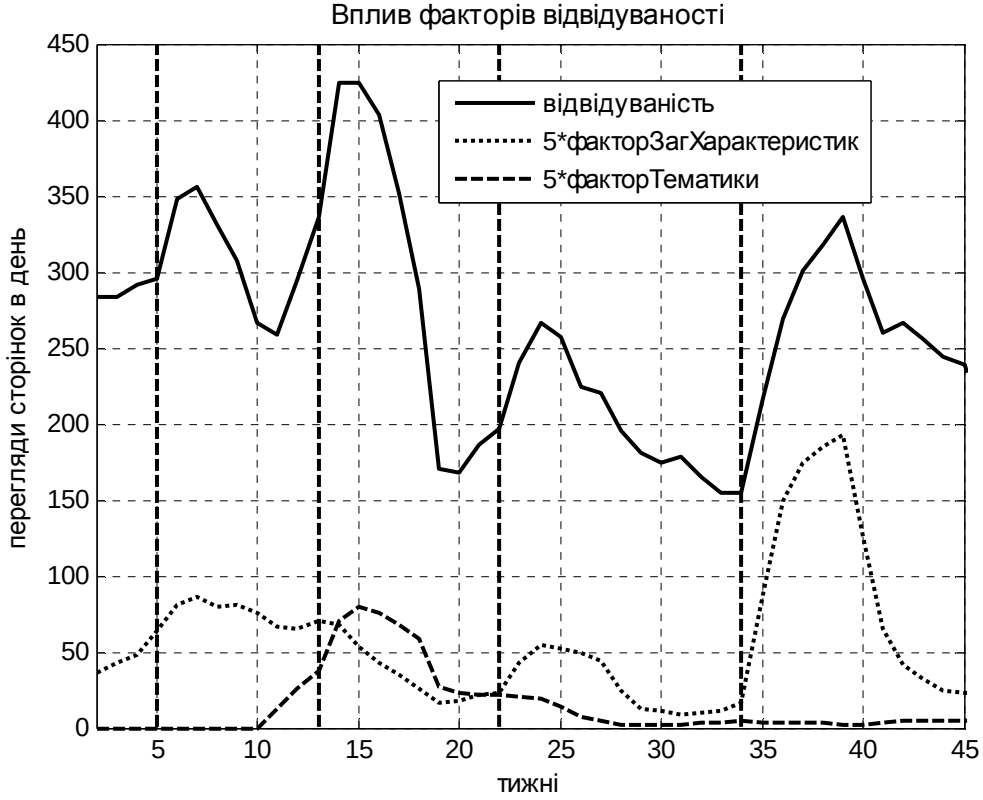

Рис. 2. Згладжена відвідуваність та фактори впливу

На основі максимальних відвідуваностей субсайтів, що активно розвиваються, виділяємо фактор впливу на відвідуваність для даного інтервалу. Аналіз згладженої динаміки факторів на стадії активації показує їх подібність, що дозволяє використовувати єдиний математичний апарат для їх моделювання. В якості такого апарату використано модифіковану систему диференціальних рівнянь Моно, які моделюють динаміку X активності фактора та потенційної його аудиторії A. Загальний вид системи Моно наводиться нижче – співвідношеннями (2)

$$\begin{cases} X'(t) = \left(a_1 \dfrac{A(t)}{a_4 + S(t)} - a_2 \right) X(t), \\ A'(t) = -a_3 \dfrac{A(t)}{a_4 + A(t)} X(t), \end{cases} \tag{2}$$

Пов'язати фактор X із обсягом згладженої відвідуваності Y достатньо точно можна за допомогою степеневої функції із відповідно підібраними параметрами – співвідношення (8).

Щоб забезпечити вихід відвідуваності на завершальній стадії на рівень d_{X0}, в правій частині диференціальних рівнянь переходимо від звичайної $A(t)$ до скоригованої відвідуваності $A_e(t)$, що визначається співвідношенням (3).

$$A_e(t) = \begin{cases} A(t) - d_{X0} \dfrac{A_0 - A(t)}{A_0 - d_{X0}} & A(t) > d_{X0}, \\ 0 & A(t) \le d_{X0}. \end{cases} \tag{3}$$

Згідно цього співвідношення скоригована відвідуваність приймає значення з інтервалу $[d_{X0}, A_0]$ на відміну від звичайної, яка теоретично може приймати значення з інтервалу $[0, A_0]$.

В результаті проведених перетворень отримаємо модель відвідуваності наступного виду:

$$\begin{cases} X'(t) = \left(a_1 \dfrac{A_e(t)}{a_4 + A_e(t)} - a_2 \right) X(t), \\ A'(t) = -a_3 \dfrac{A_e(t)}{a_4 + A_e(t)} X(t), \end{cases} \tag{4}$$

$$\text{при умові } \left(A > \frac{A_0}{2} \right) \vee \left(X \ge d_{X0} \right), \tag{5}$$

$$\begin{cases} X'(t) = 0, \\ A'(t) = 0, \end{cases} \tag{6}$$

$$\text{при умові } \left(A \le \frac{A_0}{2} \right) \wedge \left(X < d_{X0} \right), \tag{7}$$

$$Y(t) = Y(t_0) + q_1 \left(X(t) - X_0 \right)^{q_2} e^{-q_3 t} \quad , \tag{8}$$

$$X(0) = X_0, \tag{9}$$

$$A(0) = A_0 = X_{\max}, , \tag{10}$$

Ідентифікацію такої системи здійснюємо в 2 етапи. На першому етапі на основі спостереженої активності фактора X ідентифікуємо параметри $a_1 - a_4$

динаміки системи, а на другому етапі параметри моделі q_1, q_2 загальної відвідуваності. Ідентифікація здійснюється за одним із найпростіших критеріїв мінімізації середньоквадратичної похибки активності фактора X на точках навчальної вибірки. Критерій якості мінімізується за допомогою модифікованого градієнтного методу Левенберга-Марквардта, який забезпечує пошук лише локального мінімуму і тому вимагає хорошого початкового наближення. Таке початкове наближення можна побудувати за допомогою аналізу особливостей розв'язків системи диференціальних рівнянь (4). Метод ідентифікації моделі повинен бути багатоетапним, щоб врахувати можливий вплив людського фактора.

Модель (4) містить чотири невідомих параметри і тому вимагає хоча б чотирьох точок ідентифікації. Дві точки отримуємо під час встановлення початку інтервалу суттєвого приросту відвідуваності, оскільки при цьому необхідно оцінювати значення похідної відвідуваності за її різницевою апроксимацією. Доотримання ще 2-4 точок спостереження фактору впливу необхідно їх прогнозувати на основі апріорних оцінок..

Побудова оцінок такого роду вимагає хорошої інформованості про властивості шуканих функцій. Аналізуючи різні реалізації факторів впливу можна помітити, що всі вони містять по одному інтервалу зростання та спадання, однак за тривалістю ці інтервали можуть суттєво відрізнятися. Зокрема деякі із них містять лише 3-4 точки до досягнення максимальної відвідуваності. Отже процедура побудови апріорних оцінок повинна суттєво враховувати нелінійність оцінюваної функції. Це дозволяє виділити фактори впливу, які містять не більше чотирьох спостережень до досягнення максимального значення, в клас короткотермінових факторів впливу. Всі інші фактори впливу віднесемо в клас довготермінових.

Класифікувати тип факторів впливу для конкретного Веб-сайту можна, проаналізувавши динаміку його першого спостереженого фактора. Наступні прояви факторів протягом тривалого періоду будуть того ж типу.

2. Основні положення методу ідентифікації модифікованої системи Моно для короткотермінових факторів впливу

Експериментальні дослідження показують, що друга похідна короткотермінового фактора впливу має прогнозований лінійний характер. Помічена властивість дозволяє побудувати рекурентні співвідношення для покрокового прогнозу значень відвідуваності в точках, які необхідно спостерегти для ефективної ідентифікації моделі. Наближаємо значення другої похідної моделі початкового фактора впливу наступним співвідношенням :

$$q_i^0 = b_1 + b_2 i\,, \tag{11}$$

$$b_1 = \min_{t\in[t_0,t_1]}(q(t)), \quad b_3 = \max_{t\in[t_0,t_1]}(q(t)), \tag{12}$$

$$b_2 = \frac{b_3 - b_1}{t_1 - t_0}, \tag{13}$$

де q_i^0 — значення другої похідної моделі початкового фактора впливу.

Маючи оцінку параметрів наближення другої похідної будуємо багатокрокові апріорні співвідношення прогнозування значень фактору впливу.

$$\tilde{x}_1 = x_1, \tag{14}$$

$$p_0 = x_1 - x_0, \tag{15}$$

$$q_0 = b_1. \tag{16}$$

$$p_i = p_{i-1} + q_{i-1}, \tag{17}$$

$$q_i = q_{i-1} + b_2, \tag{18}$$

$$\tilde{x}_i = \tilde{x}_{i-1} + p_i. \tag{19}$$

Обчислення за співвідношеннями (15) – (17) продовжуємо поки похибка прогнозу буде лежати в допустимих межах δ_0:

$$\frac{|\tilde{x}_i - x_i|}{X_{prev}} \le \delta_0 \ , \tag{20}$$

де X_{prev} — максимальне значення попередньої реалізації фактору впливу.

Апріорні співвідношення дозволяють прогнозувати динаміку фактора впливу на попередньому етапі а також сприяють отриманню інформації про дійсні значення фактору впливу в обсягах, достатніх для застосування процедури ідентифікації моделі фактора впливу .

Ідентифікація системи Моно методом Левенберга-Марквардта вимагає побудови процедури задання початкових значень коефіцієнтів системи. При побудові початкових значень для параметрів моделі фактора впливу встановлено, що коефіцієнти a_2 і a_4 є відносно незалежними. Тому їх значення підбираємо на вузлах рівномірних сіток, які покривають деякі діапазони (21-23). Після вибору значень коефіцієнтів a_2 та a_4 визначаємо початкове значення коефіцієнтів a_1 і a_3 з рівнянь (4).

$$a_4 \in W_4 = \{a_4^0,\ a_4^0 + h_4,\ a_4^0 + 2h_4, ..., a_4^N\}. \tag{21}$$

$$a_2 \in W_2 = \{a_2^0,\ a_2^0 + h_2,\ a_2^0 + 2h_2, ..., a_2^N\}, \tag{22}$$

$$a_2^0 = \frac{p_0}{X_{\max}} \quad . \tag{23}$$

$$a_1 = \left(\frac{p_0}{x_1} + a_2\right)\frac{A_0}{A_0 + a_4} = \left(\frac{p_0}{x_1} + a_2\right)\frac{X_{\max}}{X_{\max} + a_4}. \tag{24}$$

$$a_3 = \frac{a_1}{2} \ . \tag{25}$$

Параметри сітки a_4^0, h_4, a_4^N підбираються експериментально. Для побудови мінімального значення коефіцієнта відносимо швидкість падіння інтересу аудиторії до максимальної відвідуваності Веб-сайту $X_{\max}$. Максимальне початкове значення відносної відвідуваності вибираємо кратним мінімальному, в найпростішому випадку просто подвоюючи його. Оскільки початкове значення обсягу потенційної аудиторії прирівняно до максимальної відвідуваності, то природньо покласти коефіцієнт відносного зменшення потенційної аудиторії вдвічі меншим, ніж коефіцієнт відносного приросту відвідуваності.

Для побудови початкових значень коефіцієнтів моделі відвідуваності прологарифмуємо співвідношення моделі відвідуваності Y, отримуючи систему лінійних рівнянь (26).

$$\ln(q_1) + q_2 \ln\left(X(t_i) - X_0\right) - q_3 = \ln(Y(t_i) - Y(t_0)), \quad i = 1,2,3. \tag{26}$$

Додаткове значення відвідуваності, необхідне для формування системи (26) будується на основі його лінійної екстраполяції (27) по двох перших точках .

$$\tilde{y}_3 = y_2 + (y_2 - y_1) = 2y_2 - y_1 \tag{27}$$

Ідентифікація системи Моно, а також моделі відвідуваності, здійснюється на основі мінімізації середньоквадратичних критеріїв якості:

$$\vec{a} = \arg\min_c \sum_{j=1}^{I} \left(\tilde{x}(\vec{c}, t_j) - x_j\right)^2. \tag{28}$$

$$\vec{q} = \arg\min_r \sum_{j=1}^{I} \left(\tilde{y}(\vec{r}, t_j) - y_j\right)^2. \tag{29}$$

3. Метод ідентифікації модифікованої системи Моно для довготермінових факторів впливу

Короткотермінові фактори відвідуваності характерні для Веб-сайтів, які не є життєво важливими для об'єкта, який він представляє. В іншому випадку команда, що підтримує відповідний Веб-сайт, докладає максимальних зусиль до постійного росту його відвідуваності аж до виходу на рівень, який забезпечує його належну результативність. При цьому відпрацьовуються спеціальні методики, які забезпечують досягненя максимуму відповідного приросту відвідуваності на протязі не менше шести тижнів. Для прикладу динаміки відвідуваності вибрано промо-сайт, а саме Веб-сайт сервісу API2Cart підприємства із розробки програмного забезпечення Magnetic One .

Для моделювання приросту відвідуваності даного типу використаємо раніше запропоновану модель (4)-(10). Однак ідентифікація цієї моделі в даному випадку буде мати свої особливості. Зокрема не спостерігається лінійної поведінки другої похідної фактора впливу, як у випадку короткотермінових інтервалів підвищення відвідуваності. Справа в тому, що тривалість процесу збіль-

шення обсягу фактору впливу спричиняє зміну швидкості його росту. Зокрема на першій стадії процесу росту швидкість росту відносно невелика, згодом вона суттєво зростає.

Така мінливість не дозволяє застосовувати лінійне наближення для прогнозування динаміки фактору впливу на значному інтервалі. В той же час, на початковому етапі кількість точок недостатня для застосування процедури ідентифікації. Тому на початковому етапі використовуємо співвідношення для побудови початкових значень (21)-(25) для ідентифікації коефіцієнтів системи Моно. При цьому підбір кращих значень коефіцієнтів a_2 та a_4 здійснюється за критерієм мінімізації відхилення модельного значення від спостереженого в останній на даний момент, тобто другій точці:

$$\vec{a} = \arg\min_{c} \left(\tilde{x}(\vec{c}, t_2) - x_2\right)^2. \tag{30}$$

Цього вже достатньо для ідентифікації моделі. Однак точність такої ідентифікації достатньо невисока. Тому штучно збільшуємо кількість точок ідентифікації за допомогою лінійної екстраполяції обсягів фактору впливу по двох останніх точках спостереження. Це дозволяє модифікувати критерій побудови коефіцієнтів моделі фактору наступним чином:

$$\vec{a} = \arg\min_{\vec{c}} \{ \sum_{k=1}^{m} \left(\tilde{x}(\vec{c}, t_k) - x_k\right)^2 + \\ + \sum_{k=m-1}^{m+m_1} (x_{m-1} + (x_m - x_{m-1})(k - m + 1) - \tilde{x}(\vec{c}, t_k))^2 \}, \tag{31}$$

де m — кількість спостережених точок,

m_1 — кількість додаткових точок ідентифікації.

Таким чином ідентифікована модель є достатньо точною на етапі лінійного росту фактору впливу. Однак на наступному етапі поведінка фактору впливу наближається до параболічного закону, оскільки фаза зростання відвідуваності переходить у фазу її зменшення. В цьому випадку вже похідна фактору впливу еволюціонує за законом, наближеним до лінійного. На цьому етапі функціонал похибки моделі включає як відхилення від спостережених значень, так і відхилення від лінійного наближення похідної в ще не спостережених точках:

$$\vec{a} = \arg\min_{\vec{c}} \{ \sum_{k=1}^{m_2} \left(\tilde{x}(\vec{c}, t_k) - x_k\right)^2 + \\ + \sum_{k=m-3}^{l+m_1} (p_{m-3}(m - 2 - k) + p_{m-2}(k - m + 3) - \tilde{x}'(\vec{c}, t_k))^2 \}. \tag{32}$$

3. Обчислювальні експерименти

При програмній реалізації підсистеми моделювання відвідуваності Веб-сайту використано сервіс Google Analytics та програмне середовище MatLab. В даному дослідженні аналізувалася відвідуваність Веб-сайту факультету комп'ютерних інформаційних технологій (ФКІТ- http://tanet.tneu.org/) Тернопільського національного економічного університету (ТНЕУ) майже за річний період а також відвідуваність промо сайту сервісу API2Cart компанії Magnetic One (http://www.api2cart.com/) більш ніж за річний період. Для Веб-сайту ФКІТ спостерігалися короткотермінові прирости відвідуваності, а для Веб-сайту API2Cart – довготермінові. Проведені чисельні експерименти підтверджують ефективність запропонованих методів, максимальні похибки прогнозування лежали в межах 9-13%.

Для ілюстрації особливостей застосування запропонованих методів на рисунку 3 наведено результат моделювання довготермінового фактора впливу на інтервалі лінійної екстраполяції фактора впливу.

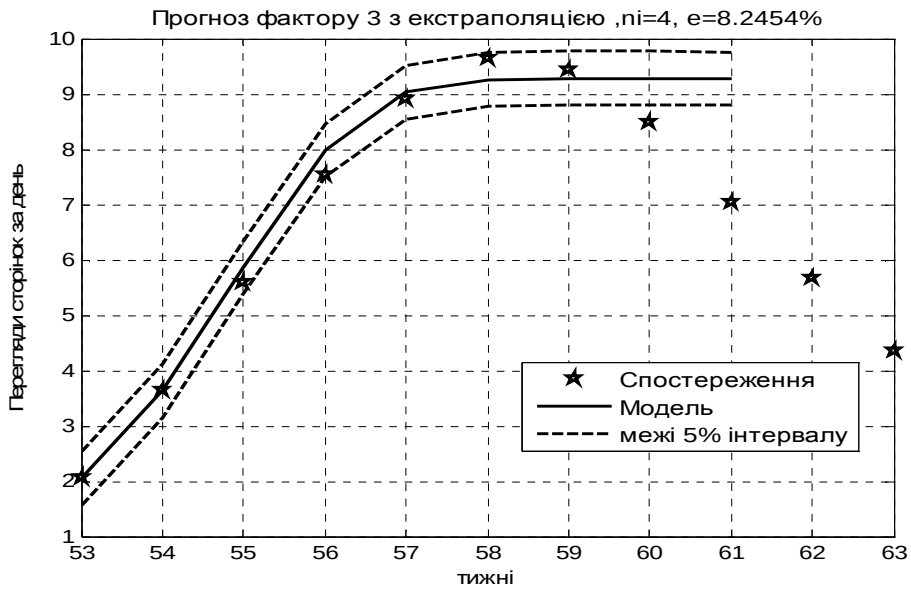

Рис.3. Результат ідентифікації моделі фактору впливу Веб-сайту API2Cart із використанням екстраполяції спостережених значень

Аналіз рисунку свідчить про задовільну точність моделі для семи спостережень. У восьмій точці спостереження точність моделі дещо перевищує 5% рівень. Тому для наступного прогнозування ідентифікуємо модель по восьми точках із застосуванням функціоналу якості, який включає окрім відхилення від спостережених значень модельованої величини відхилення похідної моделі від лінійної екстраполяції похідної фактора впливу. Результат прогнозування по моделі наводяться на рисунку 4.

Аналіз рисунку свідчить про задовільну точність прогнозування для дев'яти спостережених значень. Десяте спостереження відхиляється від прогнозу дещо більше ніж на 5%.

На відміну від моделювання фактору впливу, модель відвідуваності вдається побудувати за початковим наближенням по трьох спостережених значеннях і ця модель, побудована на основі моделі фактора впливу, не потребує подальших уточнень. Відповідний графік наведено на рисунку 5. В цьому випадку багатоетапність прогнозу відвідуваності пов'язана лише із багатоетапністю ідентифікації фактора впливу. Похибка моделі не перевищує 6.6%.

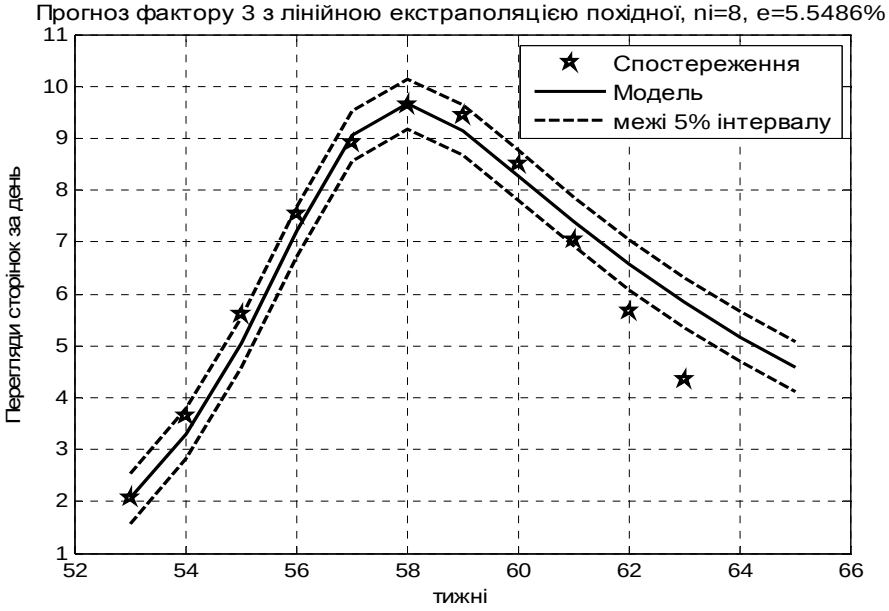

Рис.4. Результат ідентифікації моделі фактору впливу Веб-сайту API2Cart із використанням екстраполяції похідної спостережених значень

Рис.5. Результат ідентифікації моделі третього інтервалу росту відвідуваності Веб-сайту API2Cart

Висновки

У результаті проведених досліджень отримано такі наукові та практичні результати. Набула подальшого розвитку неавтономна модель відвідуваності Веб-сайтів, яка включає модель динаміки фактора впливу на основі модифікованої системи звичайних диференціальних рівнянь Моно, що на відміну від існуючих забезпечує аналітичне представлення процесів росту відвідуваності Веб-сайтів.

Вперше запропоновано метод ідентифікації моделі процесу короткотермінового росту відвідуваності Веб-сайтів, який включає етапи побудови початкового значення фактору впливу на основі апроксимації його другої похідної та процедуру побудови початкових значень коефіцієнтів моделі, щоб ідентифікувати модель із задовільною точністю.

Вперше запропоновано метод ідентифікації моделі процесу довготермінового росту відвідуваності Веб-сайтів, який використовує на окремих етапах функціонали якості, що містять лінійні екстраполяції значень фактора впливу та його першої похідної. Це забезпечує ідентифікацію моделі із задовільною точністю.

Список використаних джерел

1. Gorbunov A.L.. Markov models for website traffic// IMAT -2007, Ural University Publ., 2007, pp. 65-73
2. Khalil F. Combining. Web Data Mining Techniques for Web Page Access Prediction. // Quinsland, 2008, pp.197. http://eprints.usq.edu.au/4341/
3. Пелещишин А. М. Позиціонування сайтів у глобальному інформаційному середовищі. - Львів: Видавництво Національного університету “Львівська політехніка”, 2007.- 260с.
4. Пасічник Н.Р., Дивак М.П. Метод формування онтологічного наповнення на основі аналізу зашумленої слабкоструктурованої інформації спеціалізованих веб-сайтів // Індуктивне моделювання складних систем: Зб. наук. пр. — К.: МННЦ ІТС НАН та МОН України, 2012. — Вип. 4. — С. 158-167.
5. Пасічник Н. Метод формування онтологічного контенту, базований на аналізі інформації спеціалізованих Веб-сайтів // Вісник ХНУ: Інженерія, т.5, 2012. – С.241-244