УДК 519.766.23

І.В. БАКЛАН, А.І. ЛОГВИНЧУК, Т.В. ШУЛЬКЕВИЧ
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

КРИТЕРІЇ ПОДІБНОСТІ ЛІНГВІСТИЧНИХ МОДЕЛЕЙ

Важливою складовою лінгвістичного підходу до виявлення аномалій у часових рядах є критерій, за яким оцінюється подібність двох моделей. Наявність аномалії встановлюється шляхом порівняння моделей. Саме від вибору критеріїв залежить можливість застосування лінгвістичного підходу до аналізу часових рядів різної природи. Для вчасної реакції на ситуацію важлива швидкість виявлення аномалії. Вибір критеріїв залежить від можливості застосування мовного підходу до аналізу часових рядів різного характеру. Розглянуті основні метрики схожості текстів Геммінга, Левенштейна, Джаро-Вінклєра та ін.

Ключові слова: часові ряди, метрики текстів, лінгвістичне моделювання, лінгвістична модель.

І.В. БАКЛАН, А.І. ЛОГВИНЧУК, Т.В. ШУЛЬКЕВИЧ
Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского»

КРИТЕРИИ ПОДОБИЯ ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ

Важной составляющей лингвистического подхода к выявлению аномалий во временных рядах есть критерий, по которому оценивается сходство двух моделей. Наличие аномалии устанавливается путем сравнения моделей. Именно от выбора критериев зависит возможность применения лингвистического подхода к анализу временных рядов различной природы. Для своевременной реакции на ситуацию важна скорость выявления аномалии. Выбор критериев зависит от возможности применения языкового подхода к анализу временных рядов различного характера. Рассмотрены основные метрики сходства текстов Геммінга, Левенштейна, Джаро-Винклера и др.

Ключевые слова: временные ряды, метрики текстов, лингвистическое моделирование, лингвистическая модель.

I.V. BAKLAN, A.I. LOGVYNCHUK, T.V. SHULKEVYCH,
National Technical University of Ukraine
'The Kiev Polytechnic Institute named after Igor Sikorsky'

CRITERIA FOR SIMILARITY OF LINGUISTIC MODELS

Anomaly detection, as a form of data analytics, is finding more and more applications in various fields of human activity every year. Thus, with the growth of the IT- sector and global integration, there is an increasing need for tools to monitor cybernetic systems, respond to and remedy disruptions and failures. Intrusion detection, unauthorized access, malfunctioning in critical security systems and infrastructure management systems are among the priorities in the modern information technology world. Anomaly detection is the task of finding patterns in data that do not match expected behavior. Anomalies are inappropriate observations, emissions that are contrary to the nature of the process under study. The purpose of this work is to increase the speed of anomaly detection, in comparison with the

https://doi.org/10.32782/2618-0340/2019.2-2.2

algorithms used in automated control systems, through the use of linguistic modeling and a syntax approach to time series analysis. The presence of an anomaly is established by comparing two models - for a training dataset in which anomalies are not guaranteed and for the actual data in which we look for an anomaly. Various metrics and peer review methods are used to assess similarity. The first approach is to create a model that characterizes the normal course of the process. For this purpose, all anomalous sections are excluded from the series, or one is selected from a section of the series that does not contain deviations from the norma − thus the so-called reference row is formed. For him the construction of a linguistic model is performed. An important component of the linguistic approach to detecting anomalies in time series is the criterion by which the similarity of the two models is evaluated. The choice of criteria depends on the possibility of applying a linguistic approach to the analysis of time series of different nature. The basic metrics of the similarity of the texts by Hamming, Lowenstein, Jaro-Winkler and others are considered. During the development of the algorithm, 4 different metrics for evaluating linguistic models were considered. Given their strengths and weaknesses, root mean square was chosen as the most appropriate approach.

Keywords: time series, text metrics, linguistic modeling, linguistic model.

Постановка проблеми

Визначення аномалій як однієї з форм аналізу даних з кожним роком знаходить все більше застосувань у найрізноманітніших сферах людської діяльності. Таким чином, із зростанням сфери ІКТ та глобальної інтеграції все частіше є затребуваними інструменти для моніторингу кіберсистем, заходів реагування та усунення наслідків у разі збоїв та помилок у них. Виявлення вторгнень, несанкціонованого доступу, несправності в роботі критичних систем безпеки та систем управління інфраструктурою – одне з головних завдань у сучасному світі інформаційних технологій.

Виявлення аномалії – це задача пошуку шаблонів у даних, які не відповідають очікуваній поведінці. Аномалії – це невідповідні спостереження, викиди, що суперечать природі досліджуваного процесу.

Важливість виявлення аномалій походить від того, що аномалії даних зазвичай надають критичну інформацію про потенційну загрозу приватним, конфіденційним даним, навіть життю людини або її здоров'ю. Наприклад, ненормальна схема трафіку в комп'ютерній мережі може вказувати на те, що постраждалий комп'ютер надсилає конфіденційні дані до несанкціонованого пункту призначення. Ненормальне зображення МРТ може вказувати на злоякісну пухлину. Винятки в даних про транзакції кредитної картки можуть вказувати на крадіжку картки або посвідчення особи, а незвичні цифри датчиків зонда можуть вказувати на несправність.

Методи виявлення аномалій, що базуються на навчанні, застосовуються в багатьох прикладних областях, включаючи інформаційну безпеку, біоінформатику, автомобільну індустрію, астрономію та інші [1].

Проблема виявлення аномалій була досліджена в багатьох наукових галузях та прикладних сферах. Багато підходів та методик було спеціально розроблено для використання у певних предметних областях, тоді як деякі з ним є більш загальними.

Здебільшого підходи до виявлення аномалій є вузькоспеціалізованими та можуть застосовуватись лише в певних умовах із накладанням численних обмежень. Досі не існує такого універсального методу, який би дозволяв виконувати детекцію аномалій на довільних даних, без спостереження та з достатньою точністю [2]. Серед відомих причин можна зауважити наступні: ті дані, які є нормальними сьогодні, можуть вважатися аномалією в майбутньому (і навпаки), іноді границя між

https://doi.org/10.32782/2618-0340/2019.2-2.2

нормальними та аномальними значеннями дуже нечітка, також дуже часто існує дефіцит даних для тренування і валідації моделей.

Ще одним недоліком багатьох існуючих алгоритмів є неможливість виявляти аномалії в потокових даних реального часу, багато з яких можуть працювати лише із статистичними історичними даними, тобто аналізувати характер перебігу явищ і процесів, що вже відбулися, при достатній кількості знань про хід аналогічних процесів.

В даний час всі алгоритми виявлення аномалій можна розділити на три групи [3]: алгоритми виявлення аномалій без нагляду дослідника, виявлення аномалій із наглядом дослідника, а також комбіновані підходи.

Метою даної роботи є підвищення швидкості виявлення аномалій, в порівнянні з алгоритмами, що застосовуються в автоматизованих системах управління, за допомогою використання лінгвістичного моделювання та синтаксичного підходу до аналізу часових рядів.

Аналіз останніх досліджень і публікацій

Важливою частиною лінгвістичного підходу до виявлення аномалій у часових рядах є критерій, за допомогою яких порівнюються дві моделі. Саме від вибору критеріїв залежить можливість застосування лінгвістичного підходу до аналізу часових рядів різної природи [4–9].

Оскільки одним з основних компонентів синтаксичного методу є перетворення числових рядів до лінгвістичних послідовностей – рядків (англ. strings), доцільно припустити, що для їх порівняння можна застосовувати алгоритми визначення подібностей у текстах. До них відносяться відстань Hamming, відстань Левенштейна, схожість Jaro–Winkler та інші [10].

Алгоритм найдовшої спільної субпослідовності враховує схожість між двома рядками, що ґрунтується на основі довжини послідовностей сусідніх символів, які існують в обох рядках.

Відстань Левенштейна визначає відстань між двома рядками шляхом підраховування мінімальної кількості операцій, необхідних для перетворення одного рядка в інший, при цьому операції перетворення – це вставлення, видалення, заміна або транспозиція (перестановка) двох суміжних символів [11].

Метод Jaro визначає кількість та порядок загальних символів у двох рядках, а його модифікація – метод Jaro–Winkler також враховує спільні дворядкові префікси.

Алгоритм Needleman–Wunsch [10] є прикладом алгоритму, який використовує динамічне програмування і вперше був використаний для порівняння біологічних послідовностей (білків, амінокислот, тощо). Він виконує глобальне вирівнювання відповідної послідовності, щоб знайти найкраще співпадіння між двома послідовностями символів.

Відстань Hamming характеризує число позицій, в яких відповідні символи двох слів однакової довжини відрізняються. В більш загальному випадку, відстань Hamming використовується для рядків однакової довжини і слугує метрикою відмінності (функцією, що визначає відстань у метричному просторі) об'єктів одного розміру. Метод був запропонований Р. Геммінгом для визначення міри відмінності між кодовими комбінаціями (двійковими векторами) у векторному просторі кодових послідовностей. Hamming-відстань є окремим випадком індексу Міньковського.

Визначення Hamming distance використовується, наприклад, при кодуванні коду Грея в логічних пристроях, де є необхідність виключити так звані «логічні гонки» – відстань між суміжними кодами завжди дорівнює одиниці, тобто кожного разу змінюється тільки один біт числа. Множина слів однакової довжини створює, так

https://doi.org/10.32782/2618-0340/2019.2-2.2

званий, метричний простір, для кожної пари елементів якого визначено Hamming distance $d(x, y)$, що задовольняє аксіомам [13]:

- $d(x, y) = 0 \Leftrightarrow x = y$ (аксіома тотожності);
- $d(x, y) = d(y, x)$ (аксіома симетрії);
- $d(x, z) \leq d(x, y) + d(y, z)$ (аксіома трикутника).

Із аксіом також слідує, що відстань завжди є невід'ємною ($d(x,y)\geq 0$), а також те, що вона завжди менша за довжину слів у символах ($d(x, y)\leq n$).

Перевагою цієї метрики є простота реалізації та швидкодія – зі зростанням довжини слів час роботи алгоритму зростає лінійно.

До критичних недоліків варто віднести:

- неможливість порівняння послідовностей різної довжини – кількість елементів в обох послідовностях має бути строго однаковою;
- алгоритм враховує збіг символів лише на однакових відповідних позиціях, у той час як одні й ті ж граматичні ланцюги можуть зустрічатись у різних частинах послідовностей і при цьому не бути аномаліями.

Слід зауважити, при визначенні подібності двох лінгвістичних послідовностей даний алгоритм не доцільно застосовувати.

У теорії інформації та комп'ютерній лінгвістиці вводять метрику подібності двох лінгвістичних послідовностей (текстів). Значення цієї відстані відображається в мінімальній кількості дій вставки, заміни і видалення для перетворення одного тексту в інший. Метод названий на честь свого автора – В. Й. Левенштейна. Метод широко застосовується для виправлення помилок у словах (в пошукових системах, базах даних, при вводі тексту та автоматичному розпізнаванні відсканованого тексту або мови), порівняння текстових файлів та в біоінформатиці для порівняння генів, хромосом та білків.

В загальному випадку, операції редагування мають різну ціну:

$w(a,b)$ – ціна заміни символа a на b;

$w(\varepsilon,b)$ – ціна вставки символа b;

$w(a,\varepsilon)$ – ціна видалення символа a.

Правило трикутника застосовується, якщо дві послідовні операції можна замінити одною і це не погіршує загальну ціну (наприклад, заміна символу x на y, а потім y на z не є кращим, ніж одразу виконати заміну з x на z).

Якщо S_1 та S_2 – два рядки, що мають довжину M та N відповідно, утворені з деякого алфавіту, тоді відстань редагування Левенштейна $d(S_1, S_2)$ можна обчислити за наступною рекурентною формулою [8] $d(S_1, S_2) = D(M, N)$, де:

$$D(i,j) = \begin{cases} 0, & \text{якщо } i = 0, j = 0; \\ i, & \text{якщо } j = 0, i > 0; \\ j, & \text{якщо } i = 0, j > 0; \\ \min\begin{pmatrix} D(i,j-1)+1, \\ D(i-1,j)+1, \\ D(i-1,j-1)+m(S_1[i], S_2[j]) \end{pmatrix}, & \text{якщо } i > 0, j > 0. \end{cases}$$

Тут $m(a,\ b)$ дорівнює нулю, якщо $a = b$, та одиниці в іншому випадку; $\min(a, b, c)$ повертає найменший із аргументів, i та j – індекси символів у рядках. Пошук відстані полягає в обчисленні можливих варіантів замін, вставок і видалень та вибору найменшого значення на даному кроці.

https://doi.org/10.32782/2618-0340/2019.2-2.2

Перевагою даного методу є можливість порівнювати часові ряди у вигляді текстів, без необхідності побудови матриць передування.

Серед недоліків методу можна виділити наступні:

- потребує багато пам'яті та має відносно низьку швидкодію;
- при перестановці місцями слів або частин слів отримують відносно великі відстані;
- відстані між абсолютно різними короткими словами виявляються незначними, у той час як відстань між довгими, але дуже схожими послідовностями, є досить великою.

Враховуючи дані факти, не варто радити використання цього методу для порівняння лінгвістичних послідовностей, отриманих шляхом перетворення чисельних часових рядів, у зв'язку з їх значною довжиною.

Перевагою метода Jaro–Winkler є врахування співпадіння символів, починаючи від початку послідовності до певного моменту [13]. Це має позитивний ефект при оцінці подібності послідовностей (слів та фраз) у натуральних мовах.

Недоліком в контексті застосування відстані Jaro до порівняння часових рядів є те, що у послідовності символів однакові субпослідовності можуть знаходитись на різних позиціях. У такому випадку оцінка схожості префіксів не дає значного позитивного ефекту і відстані є завищеними, що ускладнює прийняття рішення щодо того, чи є дана послідовність аномальною.

Мета дослідження

Враховуючи недоліки вище перелічених методів подібності символьних послідовностей, постає необхідність розробити критерій порівняння двох ймовірнісних лінгвістичних моделей, що представлені у вигляді матриць переходів марковського процесу.

Викладення основного матеріалу дослідження

Є певний часовий ряд $X = \{x_1,\ x_2, x_3, \dots,\ x_m\}$, рівні якого виміряні через однакові проміжки часу. Першим кроком необхідно визначитись з алфавітом та виконати інтервалізацію ряду, кожний елемент якого характеризує певну літеру алфавіту. Оберемо в якості алфавіту множину малих латинських літер – $A = \{a, b, c, \dots, z\}$, $N = |A| = 26$. Кількість символів алфавіту визначає кількість інтервалів, по яких потрібно розподілити значення рівнів ряду. Зручно включати до алфавіту великі та малі літери, щоб використовувати їх для позначення додатних та від'ємних значень.

Для того, щоб побудувати інтервали необхідно визначити $X_{\min}$ та $X_{\max}$. Визначаємо крок інтервалу – $step = (X_{\max} - X_{\min})/N$. Розбиваємо інтервал $[X_{\min}; X_{\max}]$ на N частин:

$$[X_{min}; X_{max}] = [X_{min}; X_{min} + step], [X_{min} + step; X_{min} + 2 \times step], \dots,$$
$$[X_{min} + j \times step; X_{min} + (j+1) \times step], \dots, [X_{min} + (N-2) \times step; X_{max}].$$

Поставимо у відповідність кожному інтервалу літеру із алфавіту:

$$a = [X_{min}; X_{min} + step],$$
$$b = [X_{min} + step; X_{min} + 2 \times step],$$
$$\dots,$$
$$k = [X_{min} + j \times step; X_{min} + (j+1) \times step],$$
$$\dots,$$

https://doi.org/10.32782/2618-0340/2019.2-2.2

$$z = [X_{min} + (N - 2) \times step; X_{max}].$$

Наступний кроком виконується підстановка еквівалентних символьних значень замість числових у вихідний ряд, за наступним принципом: якщо рівень x_i потрапляє до інтервалу j $[X_{\min} + j \times step; X_{\min} + (j + 1) \times step]$, тоді замінюємо значення x_i на літеру, яка відповідає цьому інтервалу.

Таким чином, ряд $X = \{x_1,\ x_2,\ x_3, \dots ,\ x_m\}$ перетворюється на ряд $L = \{l_1, l_2, l_3, \dots, l_m\}$, де $l_i \in A$.

Виконавши перетворення з чисел у символи, потрібно побудувати матрицю передування. Для цього виконується підрахунок кількості входжень усіх символів, що передують символу, вказаному в рядку. Результат підрахунку представлено на рис. 1.

	a	*b*	*c*	...	*z*	Всього
a	14	3	12	...	2	43
b	1	3	2	...	1	15
c	5	1	4	...	3	32
...	...	...	...	...	...	22
z	0	2	3	...	5	16

Рис. 1. Кількість входжень кожного символу на відповідній позиції у послідовності.

Отримавши матрицю передувань, необхідно визначити відповідну ймовірність знаходження кожного символу на даній позицій. Для отримання матриці частот необхідно для кожного рядка розділити значення кожного елемента на суму елементів у рядку, тобто на загальну кількість символів, що зустрічаються після даного. Матриця частот зображена на рис. 2.

	a	*b*	*c*	...	*z*	Сума
a	0.32	0.069	0.27	...	0.046	1
b	0.067	0.2	0.13	...	0.067	1
c	0,156	0.031	0.125	...	0.093	1
...	...	...	...	...	...	1
z	0	0.125	0.188	...	0.313	1

Рис. 2. Частота розміщень кожного символу на відповідній позиції у послідовності.

https://doi.org/10.32782/2618-0340/2019.2-2.2

В результаті отримуємо матрицю із ймовірностями переходу станів марковського процесу, яка і є нашою лінгвістичною моделлю.

Можливі дані, у яких для побудови моделі доцільно використовувати не оригінальний ряд, а його похідні – першу, другу різниці. Використання похідних дає змогу виявити неочевидні закономірності, які важко помітити у вихідному ряді.

Робимо розрахунки для першої різниці ряду $X = \{x_1,\ x_2,\ x_3, \dots ,\ x_m\}$ є ряд $X^1 = \{x_2 - x_1, x_3 - x_2, \dots, x_m - x_{m-1}\}$, тобто ряд, який складається з попарних різниць сусідніх елементів. Заходимо другу різницю X^2, яку отримуємо шляхом застосування тих самих операцій по відношенню до членів ряду X^1.

Наявність аномалії визначається шляхом порівняння двох моделей: для навчального набору даних, у яких гарантовано відсутні аномалії, та для фактичних даних, у яких ми шукаємо аномалію. Для оцінки подібності застосовуються різноманітні метрики, а також методи експертних оцінок.

Перший підхід ґрунтується на створенні моделі, що характеризує нормальний перебіг процесу. Для цього із ряду виключаються всі аномальні ділянки та обирається одна з ділянку ряду, що не містить відхилень від норми. Таким чином утворюється, так званий, референтний ряд. Для нього виконується побудова лінгвістичної моделі.

Необхідно отримати різницю матриць P_1 та P_2, що представляють порівнювані моделі, із деякою матрицею P' третьої моделі. Наступним кроком є підрахування суми різниць всіх відповідних елементів двох матриць та порівняння отриманих для обох моделей чисел між собою. При цьому, оскільки елементами матриць є частоти (або ймовірності), з якими зустрічаються ті або інші елементи алфавіту у лінгвістичній послідовності, ми маємо справу лише з невід'ємними числами:

$$\varepsilon = \sum_{i=1}^{N}\sum_{j=1}^{N}(p_{i,j} - p'_{i,j}),$$

де N – потужність алфавіту (кількість елементів матриць);
$p_{i,j}$ – елемент матриці однієї з порівнюваних моделей P;
$p'_{i,j}$ – елемент матриці третьої моделі.

Проблемою цього підходу є те, що його результати можуть бути легко спотворені за рахунок накопичення похибки при обробці чисел з плаваючою комою. Ще одним значним недоліком є те, що даний спосіб не дає можливості розрізняти велику кількість незначних відхилень від однієї вираженої аномалії.

Більш досконалим підходом є застосування кореневого середньоквадратичного. У такому разі формула для обчислення величини різниці двох матриць (в нашому випадку матриць передування) набуває наступного вигляду:

$$\varepsilon = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(p_{i,j} - p'_{i,j})^2}.$$

Ця формула усуває основні недоліки, тобто від'ємну різницю елементів, що підносяться до квадрата. Таким чином, вплив має абсолютне значення різниці, чим більша складова різниця елементів двох матриць, тим більшу вагу вона матиме в

https://doi.org/10.32782/2618-0340/2019.2-2.2

кінцевому результаті. Операція ділення на кількість елементів матриці $N \times N$ виноситься за знак кореня і грає роль нормалізуючого коефіцієнта, зменшуючи тим самим діапазон можливих значень ε.

Висновки

Було розглянуто можливість порівняння рядів у вигляді символьних послідовностей із використанням різних метрик подібності рядків (відстань Левенштейна, відстань Hamming, схожість Jaro–Winkler), та порівнянь ймовірнісних лінгвістичих моделей.

Наведено два можливі підходи до виявлення аномалій за допомогою лінгвістичних моделей, а саме: пошук будь-яких відхилень від норми у даному часовому ряді та пошук конкретного патерну поведінки у послідовності.

Запропоновано статистичний критерій подібності ймовірнісних лінгвістичних моделей, який базується на використанні кореневого середньоквадратичного.

Список використаної літератури

1. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey. *ACM Computing Surveys*. 2009. Vol. 41. № 3. Article 15. 58 p.
2. Gupta M., Gao J., Aggarwal C. C. Han J. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*. 2014. Vol. 25. № 1. 9 p.
3. Hodge V. J., Austin J. A. Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*. 2004. Vol. 22. P. 85–126.
4. Лінгвістичне моделювання (математичне моделювання). URL: https://uk.wikipedia.org/wiki/Лінгвістичне_моделювання_(математичне_моделювання).
5. Логвинчук А. І., Баклан І. В. Застосування лінгвістичного моделювання до вирішення задачі пошуку аномалій. *Інформаційні системи та технології управління (ІСТУ2019): Матеріали ІІІ всеукраїнської науково-практичної конференції молодих вчених та студентів*. (Київ, 20-22 листопада 2019 р). Київ: НТУУ «КПІ ім. Ігоря Сікорського», 2019. С. 65–67.
6. Lohvynchuk A., Baklan I. Linguistic Approach for a Time Series Anomaly Detection. *Slovac International Scientific Journal*. 2019. Vol. 1. №35. P. 16–18.
7. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти. *Системні технології*. 2011. № 3. С. 10–19.
8. Шулькевич Т. В., Баклан І. В. Гібридний лінгвістичний підхід до моделювання часових рядів. *Прикладні питання математичного моделювання*. 2018. № 2. С. 191–202. DOI: https://doi.org/10.32782/2618-0340-2018-2-191-202
9. Баклан І. В., Шулькевич Т. В. Порівняльний аналіз прогнозу при варіації параметрів гібридної лінгвістичної моделі. *Системні технології*. 2019. Вип. 3. С. 32–41.
10. Cohen W., Ravikumar P., Fienberg S. E. A Comparison of String Distance Metrics for Name-Matching Tasks. *KDD Workshop on Data Cleaning and Object Consolidation*. 2003. Vol. 3. P. 73–78.
11. Відстань Геммінга. URL: https://uk.wikipedia.org/wiki/Відстань_Геммінга.
12. Відстань Левенштейна. URL: https://uk.wikipedia.org/wiki/Відстань_Левенштейна
13. Jaro M. Advantages in record linkage methodology as applied to the 1985 cenus of Tampa, Florida. *Journal of the American Statistical Association*. 1989. Vol. 84. Issue 406. P. 414–420.

https://doi.org/10.32782/2618-0340/2019.2-2.2

References

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*. **41**, 3, Art. 15.
2. Gupta, M., Gao, J., Aggarwal, C. C. & Han, J. (2014). Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*. **25**, 1.
3. Hodge, V. J., & Austin, J. A. (2004). Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*. **22**, 85–126.
4. Linhvistychne modeliuvannia (matematychne modeliuvannia). Retrieved from: https://uk.wikipedia.org/wiki/Лінгвістичне_моделювання_(математичне_моделювання).
5. Lohvynchuk, A. I., & Baklan, I. V. (2019). Zastosuvannia linhvistychnoho modeliuvannia do vyrishennia zadachi poshuku anomalii. Proceedings of the *Informatsiini systemy ta tekhnolohii upravlinnia (ISTU2019): Materialy III vseukrainskoi naukovo-praktychnoi konferentsii molodykh vchenykh ta studentiv*. (Kyiv, 20-22 November, 2019). Kyiv: NTUU «KPI im. Ihoria Sikorskoho», pp. 65–67.
6. Lohvynchuk, A., & Baklan, I. (2019). Linguistic Approach for a Time Series Anomaly Detection. *Slovac International Scientific Journal*. **1**, 35, 16–18.
7. Baklan, I. V. (2011). Linhvistychne modeliuvannia: osnovy, metody, deiaki prykladni aspekty. *Systemni tekhnolohii*. **3**, 10–19.
8. Shulkevych, T. V., & Baklan, I. V. (2018). Hibrydnyi linhvistychnyi pidkhid do modeliuvannia chasovykh riadiv. *Prykladni pytannia matematychnoho modeliuvannia*. **2**, 191–202. DOI: https://doi.org/10.32782/2618-0340-2018-2-191-202
9. Baklan, I. V., & Shulkevych, T. V. (2019). Porivnialnyi analiz prohnozu pry variatsii parametriv hibrydnoi linhvistychnoi modeli. *Systemni tekhnolohii*. **3**, 32–41.
10. Cohen, W., Ravikumar, P., & Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *KDD Workshop on Data Cleaning and Object Consolidation*. **3**, 73–78.
11. Vidstan Hemminha. Retrieved from: https://uk.wikipedia.org/wiki/Відстань_Геммінга.
12. Vidstan Levenshteina. Retrieved from: https://uk.wikipedia.org/wiki/Відстань_Левенштейна
13. Jaro, M. (1989). Advantages in record linkage methodology as applied to the 1985 cenus of Tampa, Florida. *Journal of the American Statistical Association*. **84**, 406, 414–420.

https://doi.org/10.32782/2618-0340/2019.2-2.2