УДК 303.722.4:[332.132:004](477)
JEL: L86

КЛАСТЕРНИЙ АНАЛІЗ ВИКОРИСТАННЯ ТА РОЗПОВСЮДЖЕННЯ ІНТЕРНЕТ-ТЕХНОЛОГІЙ У РЕГІОНАХ УКРАЇНИ

©2020 **ТУМАНОВ О. О.**

УДК 303.722.4:[332.132:004](477)
JEL: L86

Туманов О. О. Кластерний аналіз використання та розповсюдження Інтернет-технологій у регіонах України

За останні десятиліття розвиток і розповсюдження Інтернет-технологій набули величезних обертів. Використання мобільного Інтернету значно прискорило цей процес. Людям більше не потрібно залишатися вдома або в офісі, щоб перебувати в Інтернет-мережі, а деякі навіть повністю перенесли свою роботу в онлайн-середовище. Одними з важливих елементів цього середовища є соціальні мережі, блоги та інші засоби масової інформації. Соціальні медіа швидко набули популярності, оскільки дають можливість людям спілкуватися та ділитися думками. Велике значення має автоматизований аналіз даних для отримання значущої інформації, яка необхідна потенційному бізнесу, користувачам і споживачам. Для того, щоб краще вивчити використання соціальних медіа, спочатку потрібно зосередитися на загальному підході та знайти надійні показники. Ці показники можуть бути даними інформаційно-комунікаційних технологій (ІКТ), які тепер впливають на кожен аспект життя людини. Вони відіграють значну роль на робочому місці, у бізнесі, освіті та розвагах. Дана стаття включає огляд алгоритмів загальних методів кластеризації та посилання на дослідження, зроблені за останні роки, які використовували відповідні алгоритми: 1) на основі поділів; 2) на основі ієрархії; 3) на гібридній основі та 4) на основі щільності. Досліджено використання та розповсюдження Інтернет-технологій у регіонах Україні. Інформаційною базою дослідження є показники наявної ІКТ-інфраструктури в областях України у 2018 р. На основі даних використання Інтернету в регіонах України проведено кластерний аналіз та надано візуалізацію розподілів на отримані групи.

Ключові слова: *методи кластеризації, інформаційно-комунікаційні технології, аналіз, алгоритми, дослідження.*
DOI: *https://doi.org/10.32983/2222-4459-2020-3-244-252*
Рис.: *4.* ***Табл.:*** *2.* ***Бібл.:*** *17.*

Туманов Олексій Олександрович *– здобувач кафедри статистики, обліку та аудиту, Харківський національний університет ім. В. Н. Каразіна (пл. Свободи, 4, Харків, 61022, Україна)*
E-mail: *oleksii.tumanov@gmail.com*
ORCID: *http://orcid.org/0000-0003-0674-0037*

УДК 303.722.4:[332.132:004](477)
JEL: L86

Туманов А. А. Кластерный анализ использования и распространения Интернет-технологий в регионах Украины

За последние десятилетия развитие и распространение Интернет-технологий приобрели огромные обороты. Использование мобильного Интернета значительно ускорило этот процесс. Людям больше не нужно оставаться дома или в офисе, чтобы находиться в Интернет-сети, а некоторые даже полностью перенесли свою работу в онлайн-среду. Одними из важных элементов этой среды являются социальные сети, блоги и другие средства массовой информации. Социальные медиа быстро приобрели популярность, так как дают возможность людям общаться и делиться мыслями. Большое значение имеет автоматизированный анализ данных для получения значимой информации, которая необходима потенциальному бизнесу, пользователям и потребителям. Для того, чтобы лучше изучить использование социальных медиа, сначала нужно сосредоточиться на общем подходе и найти надежные показатели. Эти показатели могут быть представлены данными информационно-коммуникационных технологий (ИКТ), влияющих на все аспекты жизни человека. Они играют значительную роль на работе, в бизнесе, образовании и развлечениях. Данная статья включает обзор алгоритмов общих методов кластеризации и ссылки на исследования, сделанные за последние годы, которые использовали соответствующие алгоритмы: 1) на основе деления; 2) на основе иерархии; 3) на гибридной основе и 4) на основе плотности. Исследованы использование и распространение Интернет-технологий в регионах Украины. Информационной базой исследования являются показатели имеющейся ИКТ-инфраструктуры в областях Украины в 2018 году. На основе данных использования Интернета в регионах Украины проведен кластерный анализ и представлена визуализация распределений на полученные группы.

Ключевые слова: *методы кластеризации, информационно-коммуникационные технологии, анализ, алгоритмы, исследования.*
Рис.: *4.* ***Табл.:*** *2.* ***Библ.:*** *17.*

Туманов Алексей Александрович *– соискатель кафедры статистики, учета и аудита, Харьковский национальный университет им. В. Н. Каразина (пл. Свободы, 4, Харьков, 61022, Украина)*
E-mail: *oleksii.tumanov@gmail.com*
ORCID: *http://orcid.org/0000-0003-0674-0037*

UDC 303.722.4:[332.132:004](477)
JEL: L86

Tumanov O. O. Cluster-Analyzing the Use and Spread of Internet Technologies in the Regions of Ukraine

In recent decades, the development and spread of Internet technologies have gained enormous momentum. The use of the mobile Internet has greatly accelerated this process. People no longer need to stay at home or in an office to stay online, and some have even completely moved their work to an online environment. Social network, blogs and mass media are important elements of this environment. Social media quickly gained popularity as it enables people to communicate and share their thoughts. Automated data analysis is important to obtain meaningful information that potential businesses, users, and consumers need. In order to better learn the use of social media, the first necessity is to focus on the overall approach and find reliable indicators. These indicators can be presented by information and communication technologies (ICT) data that impact all aspects of human life. They play a significant role in work, business, education and entertainment. This article includes an overview of the algorithms of common clustering methods and references to the studies carried out in recent years that have used appropriate algorithms: 1) based on division; 2) based on hierarchy; 3) on a hybrid basis and 4) based on density. The use and spread of Internet technologies in the regions of Ukraine are researched. The information base of the research is indicators of the existing ICT infrastructure in the regions

of Ukraine as of year 2018. Based on the data on Internet use in the regions of Ukraine, a cluster analysis was conducted and visualization of distribution to the resulted groups was presented.
Keywords: *clustering methods, information and communication technologies, analysis, algorithms, research.*
Fig.: *4.* ***Tabl.:*** *2.* ***Bibl.:*** *17.*

Tumanov Oleksii O. *– Applicant of the Department of Statistics, Accounting and Auditing, V. N. Karazin Kharkiv National University (4 Svobody Square, Kharkiv, 61022, Ukraine)*
E-mail: *oleksii.tumanov@gmail.com*
ORCID: *http://orcid.org/0000-0003-0674-0037*

Узв'язку з розвитком інформаційно-комунікаційних технологій значна увага приділяється їх використанню в науково-дослідницькій сфері. Зростання кількості інформації в мережі Інтернет породжує інтерес до вивчення цієї сфери як джерела даних для наукових досліджень. За даними цифрових звітів We Are Social та Hootsuite, у 2019 р. кількість користувачів Інтернетом в Україні становить 40,91 млн осіб [14]. Соціальні медіа стають об'єктом для аналізу поведінки користувачів за допомогою різноманітних наукових методів. Одним із них є кластеризація. Методи кластеризації застосовуються для проведення територіальної диференціації, для аналізу соціальної поведінки у сферах людської діяльності, для розв'язання різноманітних завдань, таких як пристосування реклами для груп із подібними інтересами, прогнозування подій тощо.

У вітчизняній літературі дослідження з використанням кластерного аналізу в основному зосереджені на загальних соціально-економічних сферах, таких як ринок праці, сільськогосподарський сектор, фінансовий ринок тощо. Дослідників, що мають праці у цій сфері, дуже багато: Беркут О. [1], Богданова Г. [2], Єріна А. [3], Корепанов Г. [4], Корепанов О. [5], Лазебник Ю. [4], Меркулова Т. [2], Пономарьова Т. [4], Рядно О. [1], Степанов О. [5] та ін. Проте вивчення інформаційно-комунікаційного сектора, й особливо соціальних медіа, що поширені в мережі Інтернет, є недостатньо поширеним у вітчизняній літературі та потребує подальших досліджень.

Тенденція до вивчення безпосередньо соціальних медіа має широке розповсюдження в зарубіжній літературі. Так, дуже цікавими є праці таких учених, як Вісенте М. (*Vicente M.*) [10], Іфрім Дж. (*Ifrim G.*) [11], Заде Л. (*Zadeh L.*) [7], Матей К. Дж. (*Mathai K. J.*) [17], Фрідман В. (*Friedemann V.*) [13] та ін. У своїх працях дослідники використовують різні методи класифікації при дослідженні соціальної мережі Twitter на базі хештегів.

Метою даної роботи є розгляд основних методів, що використовуються в ході кластерного аналізу, та їх адаптація для виділення однорідних регіонів України за рівнем розвитку ІКТ-інфраструктури та напрямами використання мережі Інтернет.

Кластерний аналіз є важливим статистичним інструментом щодо багатовимірного аналізу даних. Він включає складні прийоми, методи та алгоритми, які можна застосовувати в різних сферах, включаючи економіку та соціальні дослідження. Метою кластерного аналізу є визначення груп подібних об'єктів відповідно до вибраних змінних. Кластерний аналіз, як правило, використовується на початку дослідження, коли дослідник не має заздалегідь обраних гіпотез чи використовуваного статистичного методу. На відміну від регресійного аналізу, для проведення якого потрібно забезпечити виконання ряду умов: вимоги нормальності, використання тільки кількісних ознак, обмеження, багатовимірний розподіл та інші, для кластерного аналізу вони не є обов'язковими [3].

Загальновживані методи кластеризації поділяються на два основні типи: а) ієрархічні та б) ітеративні (неієрархічні).

Ієрархічні методи кластеризації передбачають послідовне об'єднання елементів (об'єктів), або послідовний розподіл сукупності об'єктів. При використанні агломераційного ієрархічного методу на початку існує стільки ж кластерів, скільки об'єктів. Найбільш подібні об'єкти зливаються у групи, і ці початкові групи об'єднуються відповідно до їх подібності. У разі, якщо схожість зменшується, всі підгрупи зливаються в єдиний кластер.

Роздільні ієрархічні методи працюють у зворотному напрямку. Початкова єдина група об'єктів поділяється на дві підгрупи так, що об'єкти в одній підгрупі знаходяться далеко від об'єктів в іншій підгрупі. Кожна з цих підгруп далі також поділяється на підгрупи. Процес триває до тих пір, поки не буде стільки ж підгруп, скільки об'єктів, тобто поки кожен об'єкт не сформує окрему групу. Результати обох агломеративних методів поділу можуть бути відображені у вигляді двовимірної діаграми, відомої як дендрограма.

Дендрограма ілюструє злиття або поділи, які були зроблені на послідовних рівнях.

Дж. Іфрім (*G. Ifrim*) у своїй роботі щодо виявлення тем у Twitter використовує ієрархічну кластеризацію, спираючись на агресивну фільтрацію твітів / термінів [11].

Н. Каур (*N. Kaur*) у своєму дослідженні застосовує ієрархічний підхід для того, щоб допомогти користувачам краще розуміти твіти, групуючи їх у кластери. Мета останнього дослідження полягала в тому, щоб менша кількість кластерів була щільно сконцентрована. Робота включала використання на-

бору даних твітів, щоб побачити, як вибір функції відстані впливає на поведінку алгоритмів ієрархічної кластеризації. Для динамічного створення широких категорій подібних твітів на основі появи іменників запропоновано інтегрований ієрархічний підхід агломеративної та подільної кластеризації [16].

Наразі відомі такі типи зв'язків:

- одноразове з'єднання (мінімальна відстань, або метод найближчого сусіда);
- повне з'єднання (максимальна відстань, або метод найдальшого сусіда);
- середній зв'язок.

Також існують інші методи ієрархічної кластеризації, такі як метод Уорда та центроїдний метод.

Ітеративні методи кластеризації призначені для групування елементів не послідовно, а одночасно із урахуванням усіх обраних показників. Кількість кластерів може бути визначена заздалегідь або в ході процедури кластеризації. Оскільки непотрібно визначати заздалегідь матрицю відстані та основні дані не зберігаються під час роботи комп'ютера, то ітеративні методи можуть застосовуватися до значно більших наборів даних, ніж ієрархічні методи. Неієрархічні методи починаються або з: 1) початкового розподілу елементів на групи, або 2) з початкового набору точок, які формують ядро кластера.

Метод *k*-середніх є найбільш якісною та популярною ітеративною технікою кластеризації. Для визначеної кількості кластерів основний алгоритм передбачає виконання таких кроків:

1. Визначити параметр *k* і розділити об'єкти на *k* початкових кластерів. Число цих кластерів може бути визначене користувачем або може бути обране програмою відповідно до довільної процедури.

2. Обчислити середні або центроїди кластерів.

3. Для обраного об'єкта обчислити його відстань до кожного центроїда. Якщо об'єкт знаходиться найближче до центру власного кластера, залишити його в цьому кластері; в іншому випадку віднести його до кластера, центроїд якого найближчий до нього.

4. Повторити крок 3 для кожного випадку.

5. Повторити кроки 2, 3 і 4, поки всі об'єкти будуть у «своїх» кластерах.

В. Фрідман у своїй роботі використала метод *k*-середніх для кластеризації клієнтів компанії Nike на базі соціальної мережі Twitter [13]. За допомогою цього методу вона побудувала функції з масивного набору даних Twitter та кластеризувала їх, використовуючи міру подібності для створення угруповань користувачів.

Р. Соні та К. Дж. Матей (*R. Soni, K. J. Mathai*) [17] запропонували використання моделі «кластер – прогнозування» для поліпшення точності прогнозування настроїв у Twitter за допомогою складу навчального та контрольованого навчання. Цей алгоритм був обраний, оскільки він забезпечує задовільний компроміс між точністю, інтерпретацією та часом виконання.

Удосконаленим методом *k*-середніх є метод *Fuzzy C-means* (*FCM*) – нечіткої класифікації *c*-середніх.

Л. Заде (*L. Zadeh*) разом з іншими вченими [7] у своєму дослідженні використав метод FCM для аналізу соціальної мережі Twitter. Цей метод на основі розділів особливо підходить у випадку нечіткого групування в наборі даних. Отримані в ході дослідження нечіткі кластери були використані для отримання уявлень щодо моделей популярності хештегів та часових тенденцій. Щоб проаналізувати динаміку хештегів, автори виділили групи хештегів, які мають схожі часові уподобання, та вивчили мовні характеристики. Вони визнали найбільш і найменш репрезентативні хештеги цих груп. Прийнята методологія базується на нечіткій кластеризації, і за результатами кластерів було зроблено багато висновків щодо варіацій хештегів протягом певного періоду часу. Їх кластеризація ґрунтувалася на тому, що категоризація хештегів не є чіткою, скоріше, більшість точок даних належать до декількох кластерів відповідно до певних ступенів належності [7].

М. Вісенте (*M. Vicente*), разом зі співавторами, враховуючи лише неструктуровану інформацію, доступну для кожного твіту в профілі користувача, використав FCM-метод для гендерної класифікації користувачів [10].

Ще одним із методів кластеризації, якому варто приділити увагу, є кластеризація на основі щільності. Одним із найпоширеніших алгоритмів кластеризації на основі щільності, а також найбільш цитованим у науковій літературі є алгоритм *DBSCAN* (англ. – *density-based spatial clustering of applications with noise*). Для заданої множини точок у деякому просторі цей алгоритм відносить в одну групу точки, які розташовані найбільш щільно (точки з багатьма сусідами), та розмічає точки, які лежать в областях з невеликою щільністю (чиї сусіди розташовані занадто далеко) як викиди [12].

У дослідженні Е. Бараліс (*E. Baralis*) та інших вчених кластеризація на основі щільності була використана в контексті аналізу текстових даних Twitter, щоб виявити згуртовану інформацію, розміщену користувачами про подію, а також уявлення користувача про неї [9]. Запропонований фреймворк приймає стратегію кластеризації, яка фокусується на ділянках набору даних ітеративно та ідентифікує кластери локально. DBSCAN був використаний у ході кластерного аналізу, оскільки він дозволяє виявити кластери довільної форми, а також підвищує однорідність кластера, фільтруючи шум і викиди. Крім того, він не вимагає попереднього уточнення кількості очікуваних кластерів у даних. У цьому підході DBSCAN застосовується ітеративно на нероздільних ділянках набору даних, і всі початкові набори даних кластеризовані на першому рівні. Потім твіти, позначені як

застарілі на попередньому рівні, перегрупуються на кожному наступному рівні.

Нещодавно проведене дослідження представило застосування DBSCAN для виділення значущих сегментів твітів у пакетному режимі [8]. Сегментацію проводили на основі розрахунків оцінки «клейкості». Цей показник враховує ймовірність того, що сегмент є фразою в партії твітів (тобто локальний контекст) і ймовірність того, що він є фразою англійською мовою (тобто глобальний контекст) [15]. Потім сентиментальні варіації твітів були проаналізовані на основі цих сегментів. Кожному слову в тексті присвоювалася оцінка настрою відповідно до заздалегідь визначеної лексики настроїв. Потім почуття твіту позначається як підсумок найбільш позитивної оцінки та негативної оцінки серед окремих слів у твіті. Більше того, цей метод є менш чутливим до викидів та шуму і не вимагає початкової ідентифікації необхідної кількості кластерів. Однак для кластеризації великих обсягів даних потрібен великий об'єм пам'яті.

Як бачимо, методи кластерного аналізу дуже різноманітні та можуть задовольнити будь-які потреби дослідників. Для досягнення мети даного дослідження автором було обрано кластеризацію за методом Уорда та *k*-середніх.

За допомогою програми STATISTICA 6.0, за даними вибіркового обстеження умов життя домогосподарств України щодо доступу до Інтернету у 2018 р. було проведено кластерний аналіз регіонів України на основі наявної інформації, що стосується ІКТ-інфраструктури та напрямів використання соціальних медіа [6].

По-перше, проведемо аналіз наявної ІКТ-інфраструктури в регіонах України у 2018 р. (*рис. 1*).

У ході аналізу використовувалися такі показники:

- Var1 – кількість домогосподарств, які мають доступ до послуг Інтернету вдома;
- Var2 – населення, яке повідомило, що за останні 12 місяців користувалося послугами Інтернету;
- Var3 – абоненти рухомого (мобільного) зв'язку;
- Var4 – абоненти кабельного телебачення;
- Var5 – абоненти Інтернету.

Як бачимо, на діаграмі чітко виділяються три основні кластери.

Отже, такий попередній розподіл регіонів на групи зумовив вибір параметра $k = 3$ при застосуванні ітераційного методу *k*-середніх (*табл. 1*).

На відміну від ієрархічної процедури Уорда, ітераційна процедура оперує безпосередньо первинними даними, в ході якої формуються кластери одного рангу, ієрархічно не підпорядковані.

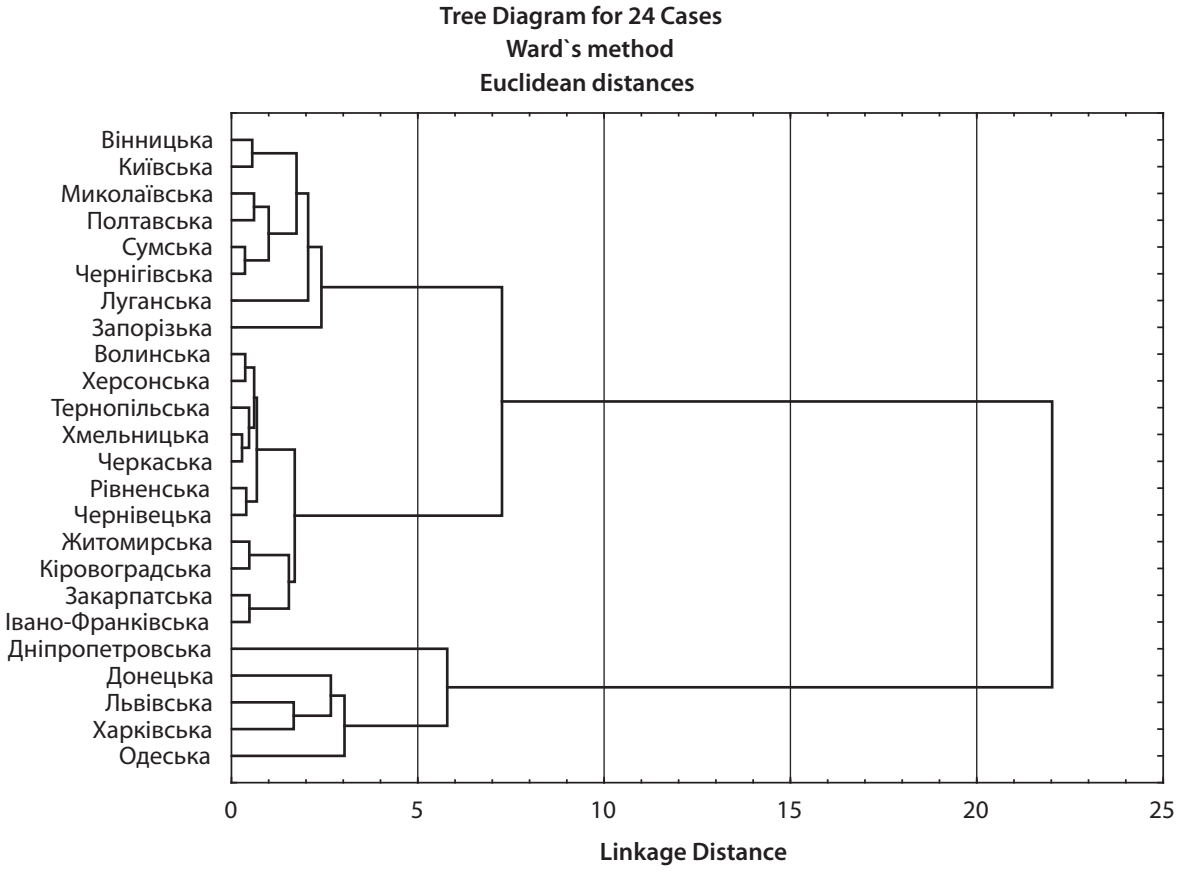

Рис. 1. Групування регіонів України за наявною ІКТ-інфраструктурою у 2018 р. методом Уорда

Джерело: авторська розробка.

Таблиця 1

Кластеризація регіонів України за наявною ІКТ-інфраструктурою у 2018 р. методом k-середніх

Номер кластера	Області України, що увійшли до кластера
1	Дніпропетровська, Донецька, Львівська, Одеська, Харківська
2	Вінницька, Запорізька, Київська, Луганська, Миколаївська, Полтавська, Сумська, Чернігівська
3	Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Рівненська, Тернопільська, Херсонська, Хмельницька, Черкаська, Чернівецька

Джерело: авторська розробка.

Алгоритм *k*-середніх реалізує ідею утворення груп за принципом «найближчого центра».

У ході реалізації методу *k*-середніх були обчислені середні нормовані значення показників для кожного з виокремлених кластерів, за якими можна впорядкувати виділені групи регіонів за рівнем наявної ІКТ-інфраструктури (*рис. 2*).

Перший кластер, до якого увійшли Дніпропетровська, Донецька, Львівська, Одеська та Харківська області, має найвищий рівень розвитку ІКТ-інфраструктури. Найнижчий рівень розвитку ІКТ-інфраструктури у 2018 р. мають області, які увійшли в третій кластер, а саме: Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Рівненська, Тернопільська, Херсонська, Хмельницька, Черкаська та Чернівецька області.

Також при розгляді графіка середніх значень змінних (див. рис. 2) привертає увагу значення змінної Var4 – «абоненти кабельного телебачення». У 1 та 2 кластерах значення цієї змінної є схожими.

Далі зосередимо увагу на аналізі користувачів Інтернету в Україні. Суспільство використовує мережу Інтернет за різними напрямами. Це може бути навчання, спілкування з друзями, отримання новин, комунікація з органами влади тощо. В ході аналізу було використано такі змінні, що стосуються розподілу населення за метою користування послугами Інтернету (у % до населення, яке повідомило, що користувалося послугами Інтернету):

- Var1 – відправлення (отримання) електронної пошти;
- Var2 – взаємодія з органами державної влади;

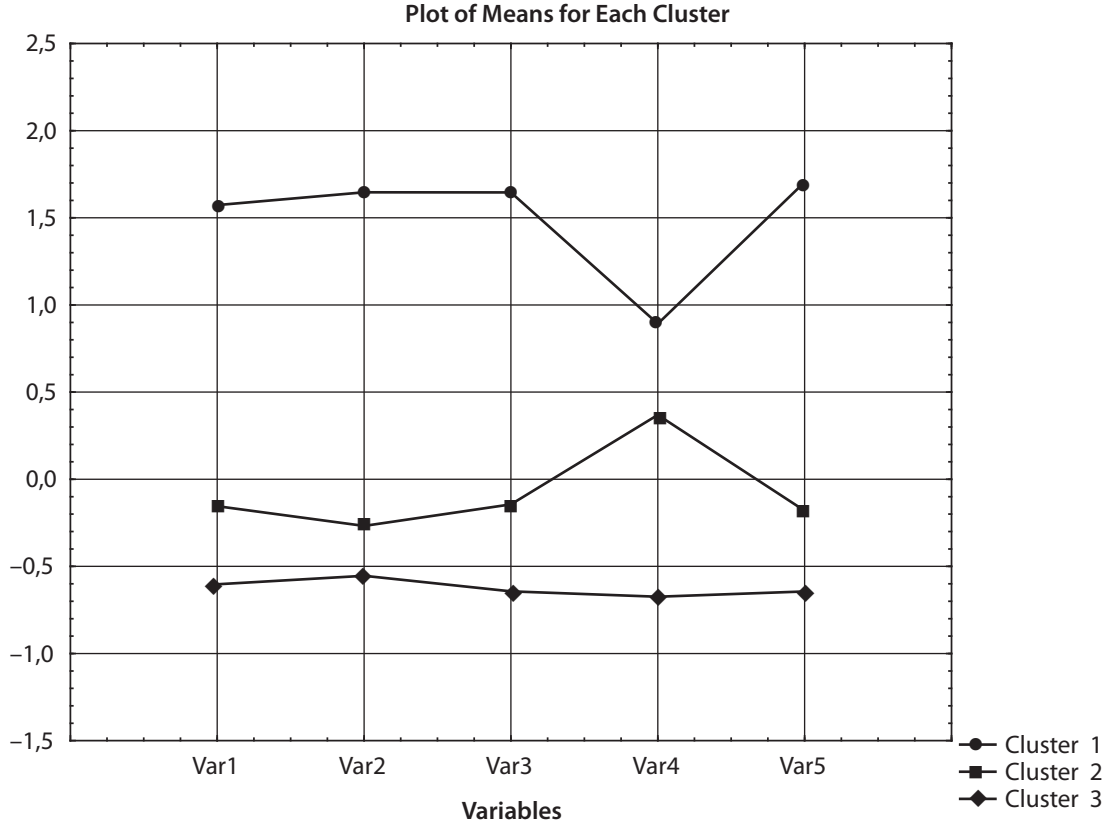

Рис. 2. Середні нормовані значення показників для кластерів наявної ІКТ- інфраструктури в Україні за 2018 р. (метод k-середніх)

Джерело: авторська розробка.

- Var3 – навчання та освіта;
- Var4 – читання/скачування газет, журналів у режимі онлайн;
- Var5 – скачування фільмів, зображень, музики; перегляд телебачення чи відео тощо;
- Var6 – гра у відео- чи комп'ютерні ігри або їх скачування;
- Var7 – скачування програмного забезпечення;
- Var8 – телефонні переговори через Інтернет (Skype, iTalk, через web-камеру);
- Var9 – спілкування (хобі);
- Var10 – банківське обслуговування;
- Var11 – пошук інформації, пов'язаної з питаннями здоров'я, як для себе, так і для інших;
- Var12 – замовлення (купівля) товарів та послуг;
- Var13 – отримання інформації щодо товарів та послуг, не названих раніше.

Дендрограму групування регіонів України за метою користування послугами Інтернету методом Уорда наведено на *рис. 3*.

Робимо припущення про доцільність розподілу регіонів України на чотири кластери з використанням методу *k*-середніх. Результати групування наведено в *табл. 2*.

Графічне представлення середніх нормованих значень показників для кожного з кластерів, отриманих у ході реалізації методу *k*-середніх, наведено на *рис. 4*.

Аналіз даних рис. 4 дозволяє зробити висновок щодо відсутності чіткого розподілу регіонів України за напрямами користування мережею Інтернет у 2018 р., які значною мірою залежать від місця користування Інтернетом. Так, під час перебування в офісі, на роботі користування Інтернетом в більшості випадків зосереджене на пошуку інформації, що стосується роботи, а не спілкування з друзями.

ВИСНОВКИ

У статті було розглянуто основні методи кластерного аналізу, що використовуються в наукових дослідженнях, а також адаптовано їх застосування до територіальної диференціації регіонів України за різними критеріями. Проведено кластеризацію регіонів із урахуванням загальної ІКТ-інфраструктури, а також напрямків користування послугами мережі Інтернет.

Встановлено, що Дніпропетровська, Донецька, Львівська, Одеська та Харківська області мали найвищий рівень розвитку ІКТ-інфраструктури у 2018 р.

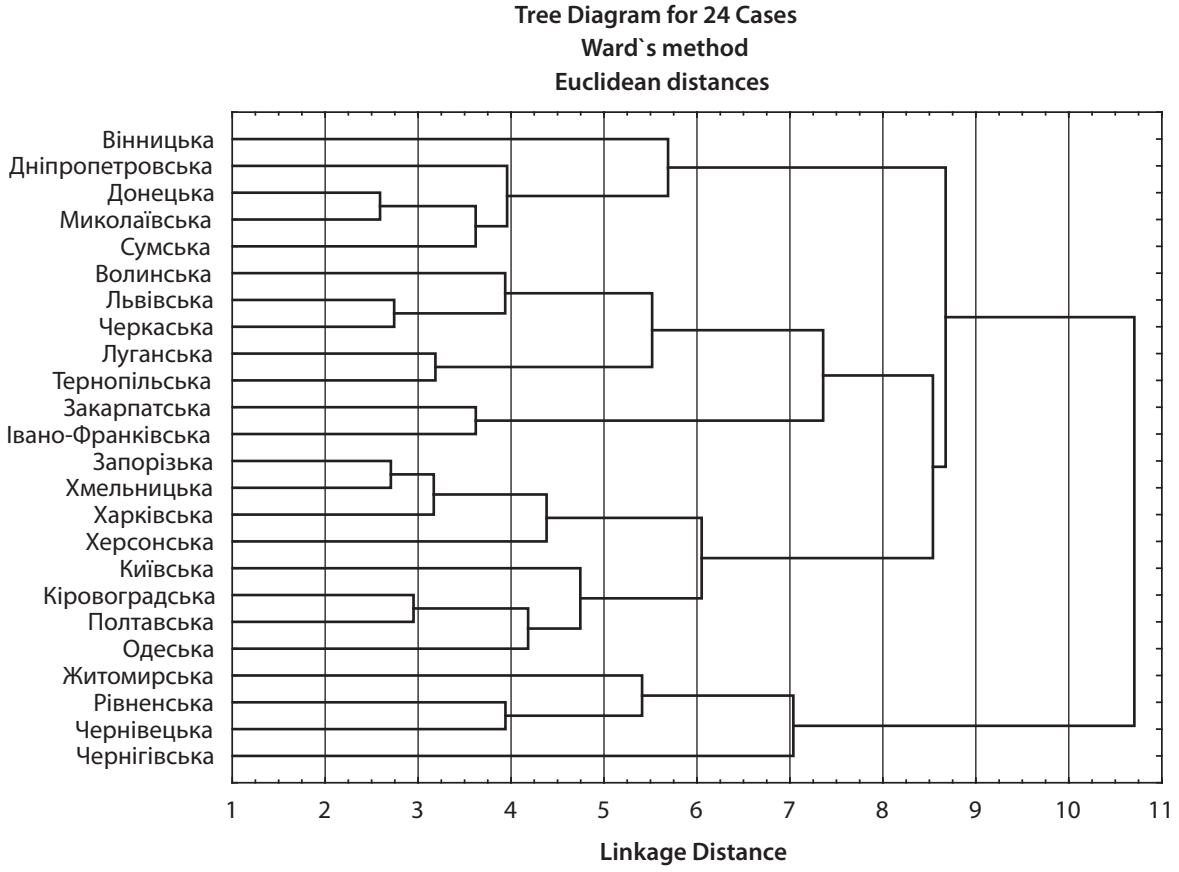

Рис. 3. Дендрограма групування регіонів України за метою користування послугами Інтернет у 2018 р. (метод Уорда)

Джерело: авторська розробка.

Таблиця 2

Кластеризація регіонів України за метою користування послугами мережі Інтернет у 2018 р. за методом k-середніх

Номер кластера	Області України, що увійшли до кластера
1	Вінницька, Житомирська, Київська, Рівненська, Чернівецька, Чернігівська
2	Запорізька, Харківська, Херсонська, Хмельницька
3	Волинська, Дніпропетровська, Донецька, Кіровоградська, Миколаївська, Одеська, Полтавська, Сумська, Черкаська
4	Закарпатська, Івано-Франківська, Луганська, Львівська, Тернопільська

Джерело: авторська розробка.

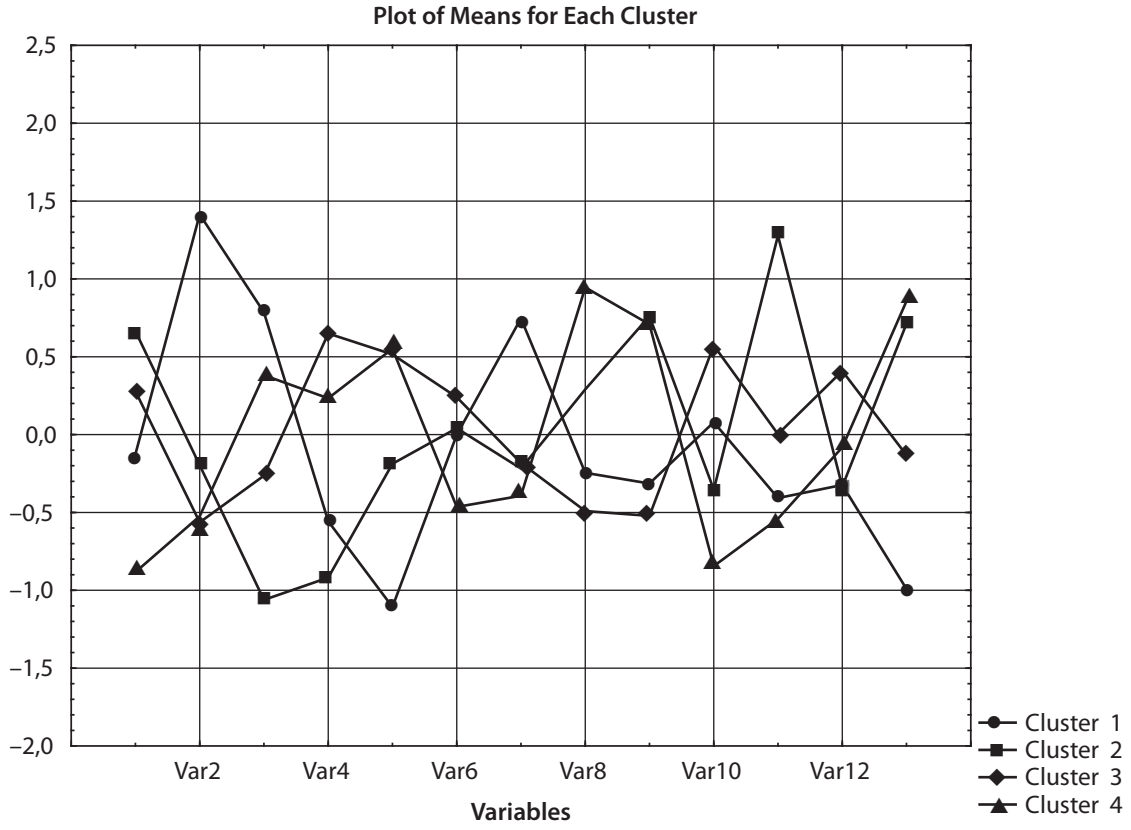

Рис. 4. Середні нормовані значення показників для кластерів за метою користування послугами мережі Інтернет в Україні у 2018 р. (метод k-середніх)

Джерело: авторська розробка.

За напрямами користування мережею Інтернет чітких відмінностей між виділеними групами регіонів України у 2018 р. не виявлено.

Дослідження використання Інтернет-послуг є актуальним і потребує подальшої роботи в цьому напрямі. ■

ЛІТЕРАТУРА

1. Рядно О. А., Беркут О. В. Дослідження структури та динаміки диференціації соціально-економічного розвитку регіонів України на основі кластерного аналізу. *Економічний вісник Донбасу*. 2016. № 1. С. 60–67. URL: https://core.ac.uk/download/pdf/87393771.pdf
2. Меркулова Т. В., Богданова Г. С. Довіра і соціально-економічний розвиток: кластерний аналіз зв'язку показників. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія «Економічна»*. 2016. Вип. 91. С. 74–79. URL: https://periodicals.karazin.ua/economy/article/view/8654/8189
3. Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. / Київ : КНЕУ, 2014, 348 с.
4. Корепанов Г. С., Лазебник Ю. О., Пономарьова Т. В. Застосування кластерного аналізу для групування регіонів за рівнем інвестиційної привабливості. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія «Економічна»*. 2014. Вип. 86. С. 63–70. URL: https://periodicals.karazin.ua/economy/article/view/5409/4956

5. Корепанов О. С., Степанов О. М. Статистичний аналіз ринку праці в Україні методами багатовимірної класифікації: регіональний аспект. *Проблеми економіки*. 2017. № 4. С. 384–392. URL: https://www.problecon.com/export_pdf/problems-of-economy-2017-4_0-pages-384_392.pdf
6. Доступ домогосподарств України до інтернету у 2018 році (за даними вибіркового обстеження умов життя домогосподарств України) : статистичний збірник. Київ : Державна служба статистики України, 2019. 45 с.
7. Zadeh L. A., Abbasov A. M., Shahbazova Sh. N. Analya sis of Twitter Hashtags: Fuzzy Clustering Approach // Fuzzy Information Processing Society (Nafips) Held Jointly with 2015 : 5th World Conference on Soft Computing (WCONSC), 2015. Annual Conference of the North American, IEEE.
DOI: 10.1109/NAFIPS-WConSC.2015.7284196
8. Anumol B., Pattani R. V. Efficient Density Based Clustering of Tweets and Sentimental Analysis Based on Segmentation. *International Journal of Computer Techniques*. 2016. Vol. 3. Issue 3. P. 53–57. URL: http://www.ijctjournal.org/Volume3/Issue3/IJCT-V3I3P9.pdf
9. Baralis E., Cerquitelli T., Chiusano S., Grimaudo L., Xiao X. Analysis of Twitter Data Using a Multiple-Level Clustering Strategy // International Conference on Model and Data Engineering. Springer, 2013. P. 13–24. URL: https://link.springer.com/chapter/10.1007/978-3-642-41366-7_2
10. Vicente M., Batista F., Carvalho J. P. Twitter Gender Classification Using User Unstructured Information // Fuzzy Systems (Fuzz-IEEE) : IEEE International Conference. IEEE, 2015. P. 1–7.
DOI: 10.1109/FUZZ-IEEE.2015.7338102
11. Ifrim G., Shi B., Brigadir I. Event Detection in Twitter Using Aggressive Filtering and Hierarchical Tweet Clustering // Second Workshop on Social News on the Web (Snow). Seoul, Korea, 8 April 2014, ACM. URL: http://ceur-ws.org/Vol-1150/ifrim.pdf
12. DBSCAN // Вікіпедія. URL: https://uk.wikipedia.org/wiki/DBSCAN
13. Friedemann V. Clustering A Customer Base Using Twitter Data. 2015. URL: https://pdfs.semanticscholar.org/08cd/1743d71b9f3e54208871c1562c6083b25f24.pdf
14. Global Digital Report 2019 – We are Social. URL: https://wearesocial.com/global-digital-report-2019
15. Li C., Sun A., Weng J., He Q. Tweet Segmentation and Its Application to Named Entity Recognition. 2015.
DOI: 10.1109/TKDE.2014.2327042
16. Kaur N. A Combinatorial Tweet Clustering Methodology Utilizing Inter and Intra Cosine Similarity. Regina, 2015. URL: https://ourspace.uregina.ca/bitstream/handle/10294/6549/Kaur_Navneet_200331665_MASC_SSE_Fall2015.pdf?sequence=1
17. Soni R., Mathai K. J. Improved Twitter Sentiment Prediction Through Cluster-Then-Predict Model. International Journal of Computer Science and Network. 2015. Vol. 4. Issue 4. P. 559–563. URL: https://arxiv.org/ftp/arxiv/papers/1509/1509.02437.pdf

Науковий керівник – Чала Т. Г., кандидат економічних наук, доцент кафедри статистики, обліку та аудиту Харківського національного університету ім. В. Н. Каразіна

REFERENCES

Anumol, B., and Pattani, R. V. "Efficient Density Based Clustering of Tweets and Sentimental Analysis Based on Segmentation". International Journal of Computer Techniques. 2016. http://www.ijctjournal.org/Volume3/Issue3/IJCT-V3I3P9.pdf

Baralis, E. et al. "Analysis of Twitter Data Using a Multiple-Level Clustering Strategy". International Conference on Model and Data Engineering. Springer, 2013. https://link.springer.com/chapter/ 10.1007/978-3-642-41366-7_2

"DBSCAN". Vikipediia. https://uk.wikipedia.org/wiki/DBSCAN

Dostup domohospodarstv Ukrainy do internetu u 2018 rotsi (za danymy vybirkovoho obstezhennia umov zhyttia domohospodarstv Ukrainy) : statystychnyi zbirnyk [Access of Households of Ukraine to the Internet in 2018 (According to a Sample Survey of Living Conditions of Households in Ukraine): A Statistical Collection]. Kyiv: Derzhavna sluzhba statystyky Ukrainy, 2019.

Friedemann, V. "Clustering A Customer Base Using Twitter Data". 2015. https://pdfs.semanticscholar.org/08cd/1743d71b9f3e54208871c1562c6083b25f24.pdf

"Global Digital Report 2019 - We are Social". https://wearesocial.com/global-digital-report-2019

Ifrim, G., Shi, B., and Brigadir, I. "Event Detection in Twitter Using Aggressive Filtering and Hierarchical Tweet Clustering". Second Workshop on Social News on the Web (Snow). 2014. http://ceur-ws.org/Vol-1150/ifrim.pdf

Kaur, N. "A Combinatorial Tweet Clustering Methodology Utilizing Inter and Intra Cosine Similarity". Regina, 2015. https://ourspace.uregina.ca/bitstream/handle/10294/6549/Kaur_Navneet_200331665_MASC_SSE_Fall2015.pdf?sequence=1

Korepanov, H. S., Lazebnyk, Yu. O., and Ponomaryova, T. V. "Zastosuvannia klasternoho analizu dlia hrupuvannia rehioniv za rivnem investytsiinoi pryvablyvosti" [Using Cluster Analysis to Regions Grouping by the Degree of Investment Appeal]. Visnyk Kharkivskoho natsionalnoho universytetu imeni V. N. Karazina. Seriia «Ekonomichna». 2014. https://periodicals.karazin.ua/economy/article/view/5409/4956

Korepanov, O. S., and Stepanov, O. M. "Statystychnyi analiz rynku pratsi v Ukraini metodamy bahatovymirnoi klasyfikatsii: rehionalnyi aspekt" [Statistical Analysis of the labor Market in Ukraine Using Multidimensional Classification Methods: the Regional Aspect]. Problemy ekonomiky. 2017. https://www.problecon.com/export_pdf/problems-of-economy-2017-4_0-pages-384_392.pdf

Li, C. et al. *Tweet Segmentation and Its Application to Named Entity Recognition*, 2015.
DOI: 10.1109/TKDE.2014.2327042

Merkulova, T. B., and Bohdanova, H. C. "Dovira i sotsialno-ekonomichnyi rozvytok: klasternyi analiz zviazku pokaznykiv" [Trust and Socio-Economic Development: Cluster Analysis of Parameter Interdependencies]. Visnyk Kharkivskoho natsionalnoho universytetu imeni V. N. Karazina. Seriia «Ekonomichna». 2016. https://periodicals.karazin.ua/economy/article/view/8654/8189

Riadno, O. A., and Berkut, O. V. "Doslidzhennia struktury ta dynamiky dyferentsiatsii sotsialno-ekonomichno-

ho rozvytku rehioniv Ukrainy na osnovi klasternoho analizu" [A Study of the Structure and Dynamics of Differentiation of Social and Economic Development of Ukraine Based on a Cluster Analysis]. Ekonomichnyi visnyk Donbasu. 2016. https://core.ac.uk/download/pdf/87393771.pdf

Soni, R., and Mathai, K. J. "Improved Twitter Sentiment Prediction Through Cluster-Then-Predict Model". International Journal of Computer Science and Network. 2015. https://arxiv.org/ftp/arxiv/papers/1509/1509.02437.pdf

Vicente, M., Batista, F., and Carvalho, J. P. "Twitter Gender Classification Using User Unstructured Information". Fuzzy Systems (Fuzz-IEEE) : IEEE International Conference. IEEE, 2015. 1-7.
DOI: 10.1109/FUZZ-IEEE.2015.7338102

Yerina, A. M. *Statystychne modeliuvannia ta prohnozuvannia* [Statistical Modeling and Forecasting]. Kyiv: KNEU, 2014.

Zadeh, L. A., Abbasov, A. M., and Shahbazova, Sh. N. "Analysis of Twitter Hashtags: Fuzzy Clustering Approach". Fuzzy Information Processing Society (Nafips) Held Jointly with 2015 : 5th World Conference on Soft Computing (WCONSC). IEEE, 2015.
DOI: 10.1109/NAFIPS-WConSC.2015.7284196