DOI: 10.35681/1560-9189.2021.23.2.239209

УДК 004.67

Д. В. Ланде[1], Л. Страшной[2], І. В. Балагура[1]
[1]Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна
[2]Каліфорнійський університет в Лос-Анджелесі
Хілгард Авеню, 405, Лос-анджелес, США

Метод формування та кластеризації кореляційних мереж понять

Для вирішення задачі формування та кластеризації понять запропоновано методику формування, кластеризації, ранжирування та подальшої візуалізації спрямованих кореляційних мереж, зв'язки яких визначаються на основі рядів динаміки, що відповідають цим поняттям. Як приклади розглянуто часові ряди динаміки вживання термінів, що формуються сервісом Google Books Ngram Viewer для формування кореляційної мережі наукових понять, і часові ряди динаміки захворюваності на коронавірус у різних країнах для формування та кластерізації мережі країн за ознакою подібності відповідних статистичних рядів. Наведена методика може застосовуватися з метою узагальнення множини сутностей без явно виражених зв'язків між ними на основі даних, отриманих в аналітичних системах різного призначення.

Ключові слова: *кореляційна мережа, динаміка публікацій, Google Books Ngram Viewer, візуалізація мережевих структур, кластерний аналіз.*

Вступ

Мережеві структури набули широкого застосування, вони використовуються в пошукових системах, у мапах, для захисту комп'ютерних мереж, для дослідження взаємодії груп, у соціальних мережах. Окремий клас серед них — кореляційні мережі, які формуються на основі обчислення попарних кореляцій між змінними, застосовуються в багатьох галузях знань, таких як прогнозування клімату, фінансовий маркетинг і біоінформатика [1]. Сучасні інформаційні технології неможливо уявити без методів і засобів обробки мережевих структур, але не завжди структурні особливості виражені явно. Завжди виникає питання, як побудувати мережу, щоб застосувати широкий спектр методів і засобів її обробки, отримати й інтерпретувати результати, якщо в розпорядженні у дослідника є лише деякі сутності — вузли, але явно не визначені зв'язки між ними. Якщо кожну сутність можна представити у вигляді однорідного багатовимірного вектора параметрів, можливо налагодити зв'язки подібності, застосувати методи класифікації або кластерного аналізу для

виявлення груп подібних документів. Задача кластеризації об'єктів широко використовується для дослідження в різних галузях інженерії, медицини, науки та в повсякденному житті [2].

У представленій роботі пропонується метод, який ставить у відповідність сутності (поняттю з предметної області) вектор динаміки, відповідний розподілу документів за датами (роками). Більш конкретно, кожному року ставиться у відповідність число — кількість появ суті в публікаціях, які охоплюються, наприклад системою Google Books. Розмірність цього вектора відповідає кількості років, довжині інтервалу часу, протягом якого аналізувався масив публікацій.

Метод формування та кластеризації кореляційних мереж

Для побудови мережевих структур для кожного поняття/сутності формуються вектори, що відповідають цим сутностям, наприклад динаміка публікацій за роками. Після формування векторів, які відповідають окремим сутностям, формується кореляційна мережа, яку можна розглядати як спосіб зберігання та візуалізації сутностей, які об'єктивно пов'язані між собою [3]. Дійсно, можна сформувати вектори динаміки для різних сутностей, зв'язок між якими не завжди є явним.

Нижче пропонується метод побудови мережі взаємозв'язку сутностей (понять), що складається з наступних етапів [4, 5]. Загальну схему методу наведено на рис. 1.

1. Для кожної сутності формується запит до сервісу Google Books Ngram Viewer. Також визначається період аналізу — розмірність відповідних векторів динаміки.

2. У результаті виконання запитів визначається безліч векторів динаміки, що відповідають наведеним поняттям;

3. Обчислюється безліч максимальних крос-кореляцій між отриманими векторами, формується відповідна кореляційна матриця з елементами:

$$a_{ij} = \max_m \frac{\sum_{k=1}^{n-m} w_{k+m}^i w_k^j}{\sqrt{\sum_{k=m+1}^{n} \left(w_k^i\right)^2} \sqrt{\sum_{k=1}^{n-m} \left(w_k^j\right)^2}}. \quad (1)$$

Тут кожній сутності s_k із множини $S = \{s_k\}_{k=1}^{|S|}$ ставиться у відповідність вектор значень параметрів $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, де n — кількість елементів у множині параметрів. Функція max використовується з тих міркувань, що близькі за своєю природою процеси можуть мати близьку за динамікою поведінку, але можливо із зсувом за часом.

4. Здійснюється формування матриці суміжності відповідно до формули (1) і збереження цієї матриці у файлі у форматі CSV. У зв'язку з тим, що в таблиці суміжності існують зв'язки між усіма вузлами, відповідно до [6], ігноруються зв'язки, значення яких менше деякого вибраного порога. Вибір цього порога повністю залежить від досвіду аналітиків. В інформаційній технології, яка описується, формується кореляційна матриця і передається для обробки і візуалізації системі аналізу мережевих структур Gephi (https://gephi.org/) [7]. Gephi — це найпоширеніша програма візуалізації і аналізу мережевих структур, яка забезпечує швидку компонов-

ку, ефективне дослідження даних, а також візуалізацію великомасштабних мереж. Разом з цим, матриця суміжності у форматі CSV для системи Gephi має деякі особливості, які необхідно враховувати (нулі по діагоналі, розташування символів «;» та інші).

5. Здійснюється завантаження значень цієї матриці у форматі CSV у систему Gephi. Ця система має ряд режимів, серед яких для отримання мережевих характеристик застосовується режим «Лабораторія даних». У цьому режимі, крім звичайних ступенів вузлів матриці, можна розрахувати їхні значення за алгоритмами PageRank, Hits, модулярністю тощо. Крім того, існують можливості ранжирування вузлів матриці (сутностей) за цими параметрами.

6. Здійснюється визначення класів модулярних груп об'єктів і подальша кластеризація завантаженої мережевої структури [6, 7]. Модулярність обчислюється як різниця між часткою ребер всередині кластера в розглянутій мережі і очікуваною часткою ребер всередині кластера в мережі, в якій вершини мають ту ж ступінь, і у вихідній, але ребра розподілені випадково. Модулярність мережі можна виразити формулою:

$$Q = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - \frac{k_i k_j}{2m}\right]\ \delta\left(c_i, c_j\right), \tag{2}$$

де a_{ij} — елемент матриці суміжності A; m — кількість ребер у графі; k_i, k_j — ступені відповідних вузлів; δ — дельта функція Кронекера (показує чи знаходяться вузли i та j в одному класі модулярності).

7. Візуалізація мережі виконується в системі Gephi.

8. На останньому етапі виконується експертна інтерпретація результатів.

Удосконалення методу

Пропонується враховувати два моменти при побудові кореляційної мережі, а саме враховувати:

1) який процес почався першим;

2) абсолютні значення часових рядів при взаємній кореляції, тобто значення направленого зв'язку між вузлами A та B визначати пропорційно сумі значень числового ряду, відповідного вузла A.

Нехай кожному елементу s_k із множини об'єктів $S = \left\{s_k\right\}_{k=1}^{|S|}$ ставиться у відповідність вектор значень параметрів $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, де n — кількість елементів у цій множині.

Для реалізації пункту 1 формула визначення зв'язку між об'єктами i та j (1) буде мати вигляд:

$$a_{ij} = \max_{0<m\le K} \frac{\sum_{k=1}^{n-m} w_{k+m}^i w_k^j}{\sqrt{\sum_{k=m+1}^{n}\left(w_k^i\right)^2}\sqrt{\sum_{k=1}^{n-m}\left(w_k^j\right)^2}}, \tag{3}$$

де K — ширина вікна можливих зсувів по часу.

Функція max використовується тому, що близькі за природою процеси можуть мати близьку за динамікою природу, але можливо із зсувом за часом. На відміну від описаного вище методу, врахування m виконується не по спектру значень $[-K, K]$, а в інтервалі $[1, K]$.

Для реалізації другого пункту, кожен із елементів матриці a_{ij} домножується на значення суми значень відповідного вектора $v_i = C\sum_{k=1}^{n} w_k^i$, де C — нормуюча константа.

При подальшому використанні засобів візуалізації Gephi мережа визначається як ненаправлена, розміри вузлів відповідають ступеням вузлів зваженої спрямованої мережі. Кластеризація, у разі необхідності, обчислюється за алгоритмами OpenOrd або Fruchterman Reingold, а модулярність вузлів розраховується з Resolution = 0.5.

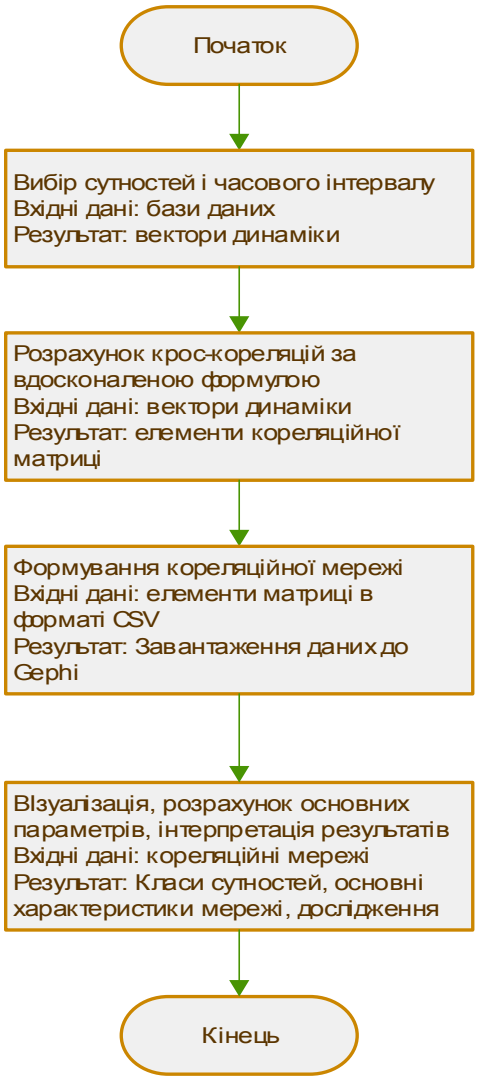

Рис. 1. Метод формування та кластеризації кореляційних мереж

Приклади

Як демонстраційний приклад розглянемо три сутності (Node1, Node2, Node3), кожному з яких відповідає часовий ряд:

Node1: (0, 1, 2, 3, 4, 5, 4, 3, 2, 1, 0, 0, 0);
Node2: (0, 0, 0, 0, 1, 2, 3, 4, 3, 2, 1, 0, 0);
Node3: (0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0).

Кореляційна матриця, що відповідає прикладу, містить наступні елементи:

Node1;Node2;Node3;
Node1;0.000;0.818;0.623;
Node2;0.818;0.000;0.766;
Node3;0.623;0.766;0.000.

Кореляційна мережа наведена на рис. 2.

У цій матриці вузол 2 представлений самим великим колом, хоча очевидно, що процес, відповідний вузлу 1, почався раніше і має велику амплітуду.

Виправити цю невідповідність дозволяє представлений удосконалений алгоритм, у результаті якого отримуємо наступну матрицю зв'язків вузлів, візуалізація якої приведена на рис. 3:

Node1;Node2;Node3;
Node1;0.000;1.022;0.779;
Node2;0.611;0.000;0.613;
Node3;0.002;0.050;0.000.

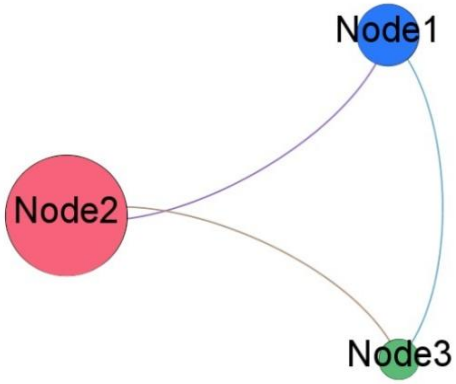

Рис. 2. Кореляційна мережа, що відповідає прикладу, розрахована за алгоритмом із [5]

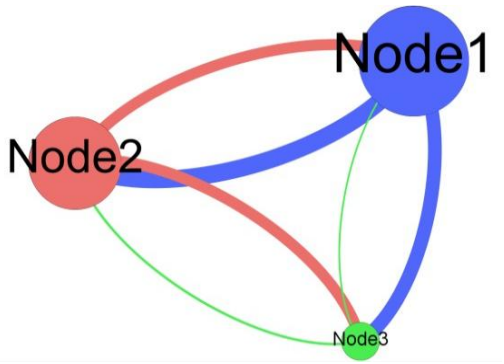

Рис. 3. Мережа понять на основі удосконаленого методу

Приклад формування та кластеризації кореляційних мереж понять на основі динаміки публікацій

Джерелом даних у наведеному прикладі використовується сервіс Google Books Ngram Viewer (https://books.google.com/ngrams). Даний сервіс дозволяє, зокрема, отримувати масиви чисел, що відповідають відносній частоті появи термінів у публікаціях за роками. Для отримання цих масивів, до системи можна звернутися через інтерфейс користувача шляхом введення запиту — назви терміну.

На рис. 4 наведено фрагмент інтерфейсу отримання динаміки відповідної тематики «Штучний інтелект» (Artificial Inteligence).

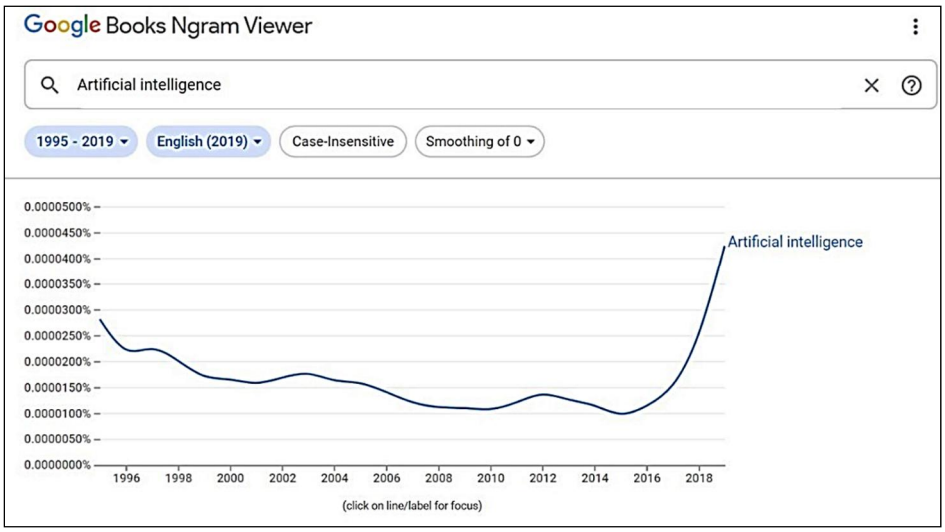

Рис. 4. Фрагмент інтерфейсу сервісу Google Books Ngram Viewer, на якому у вигляді графіка представлений вектор динаміки появи поняття Artificial Inteligence

Для побудови мережі понять, пов'язаних сучасними напрямками Computer Science, за джерела інформації у роботі розглядалися дані, які отримані шляхом звернення до сервісу Google Books Ngram Viewer. Як приклад розглядається 20 понять, перерахованих у таблиці. Також визначається період аналізу (1995–2019 рр.).

Терміни-запити до сервісу Google Books Ngram Viewer

N	Сутність	Скорочення
1	Big data	BDT
2	Complex networks	CNT
3	Machine learning	MNL
4	Deep learning	DPL
5	Neural networks	NNT
6	Data mining	DTM
7	Semantic web	SWB
8	Pattern Recognition	PTR
9	Complex systems	CST
10	Artificial intelligence	ARI
11	Smart grids	SMG
12	Social computing	SCC
13	Natural language processing	NLP
14	Informetrics	INM
15	Social network analysis	SNA
16	Information retrieval	INR
17	Information extraction	INE
18	Computer vision	CMV
19	Digital libraries	DLB
20	Recommender Systems	RSS

За допомогою сервісу Google Books Ngram Viewer визначаються вектори динаміки, що відповідають заданим поняттям, представлені у форматі JSON у вихідному коді вихідної форми (рис. 5).

```
  ngrams.data = [{"timeseries": [9.575330750521971e-09, 6.9888743681190135e-09,
1.2791656622823666e-08, 1.1356319440380958e-08, 1.2087312484254653e-08,
1.0711201703372808e-08, 1.2456910170044466e-08, 3.029423822908939e-08, 1.1776428721077536e-
08, 7.028031934197543e-09, 8.192412970231544e-09, 6.390932227873236e-09,
8.478197699446355e-09, 7.651633993077667e-09, 6.700836774342633e-09, 5.9736504631757725e-
09, 4.666732333902246e-09, 1.3329707115872225e-08, 8.425238284814895e-09,
1.3226189032877755e-08, 1.0391362437189855e-08, 4.0528647105020355e-08, 2.276777522070006e-
08, 1.8817276625782142e-08, 1.8356137942987516e-08], "parent": "", "ngram": "Informetrics",
"type": "NGRAM"}, {"timeseries": [6.626123649766669e-08, 4.4549572919549973e-08,
4.416813581542556e-08, 3.641325463377143e-08, 3.4242845003973343e-08, 3.139854243272566e-
08, 2.8454564926505554e-08, 2.8115911376858094e-08, 3.27441078695756e-08,
3.3102121932415685e-08, 3.2795835380738936e-08, 3.006990922926889e-08, 2.393094433728038e-
08, 2.0940479572573167e-08, 2.411564814508438e-08, 2.3531855575242844e-08,
3.441038387563822e-08, 5.1720220994866395e-08, 3.2820665296640007e-08, 2.867479231838388e-
08, 3.108477386604136e-08, 3.396015557655119e-08, 4.183678115055045e-08,
4.8998536783528834e-08, 7.350192987587434e-08], "parent": "", "ngram": "Computer vision",
"type": "NGRAM"}, {"timeseries": [2.1733773891696728e-08, 2.3040740870783338e-08,
2.617509764490933e-08, 2.3740966526020202e-08, 1.967496743304764e-08, 1.9771428938497593e-
08, 1.961558915297701e-08, 1.848783703906065e-08, 1.9918157789788893e-08,
2.0516173648843505e-08, 1.9428796349529875e-08, 2.2530743493121008e-08,
1.8192798378890984e-08, 1.2451295106075122e-08, 1.1564347701664701e-08,
1.6192725382779827e-08, 1.944471783588142e-08, 1.9834835995879985e-08, 1.9200374623551397e-
08, 2.3690152062272318e-08, 2.5081311250119143e-08, 2.88420132221745e-08,
3.207855669984383e-08, 4.345080029111159e-08, 6.216553316562567e-08], "parent": "",
"ngram": "Natural language processing", "type": "NGRAM"}];
```


Рис. 5. Вихідний код

У результаті аналізу 20-ти понять було отримано відповідну направлену кореляційну матрицю, сформовано мережу та проведено її кластеризацію (рис. 6).

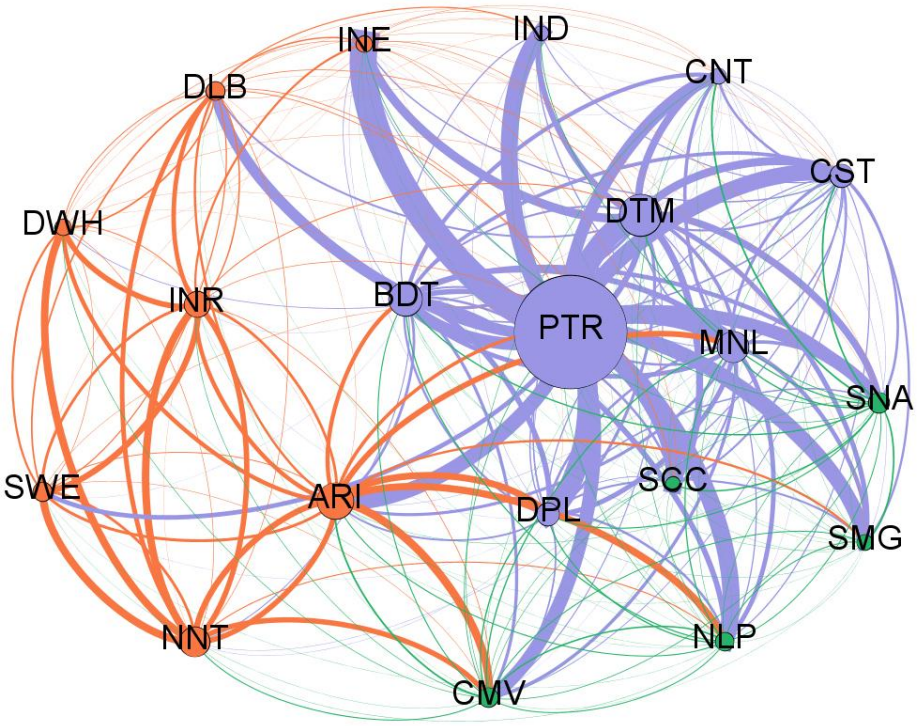

Рис. 6. Мережа сутностей (понять) у середовищі системи Gephi

На рис. 7–9 наведено типові динаміки понять (сутностей), що входять у різні кластери.

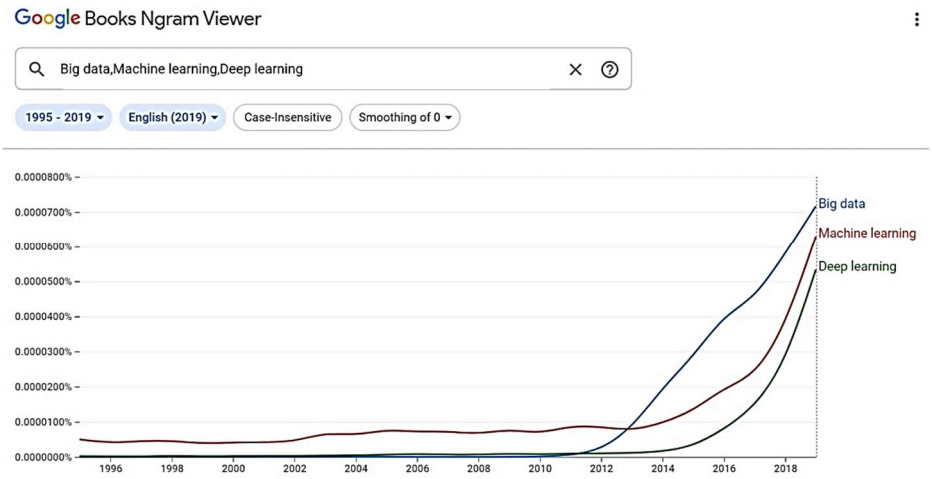

Рис. 7. Динаміка сутностей (кластер Big data, Machine learning, Deep learning)

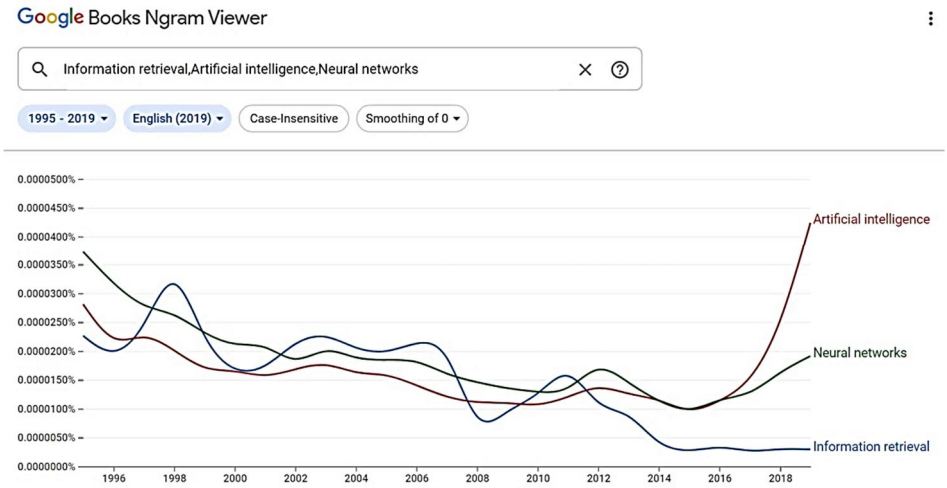

Рис. 8. Динаміка сутностей (кластер Artificial intelligence, Neural networks, Information retrieval)

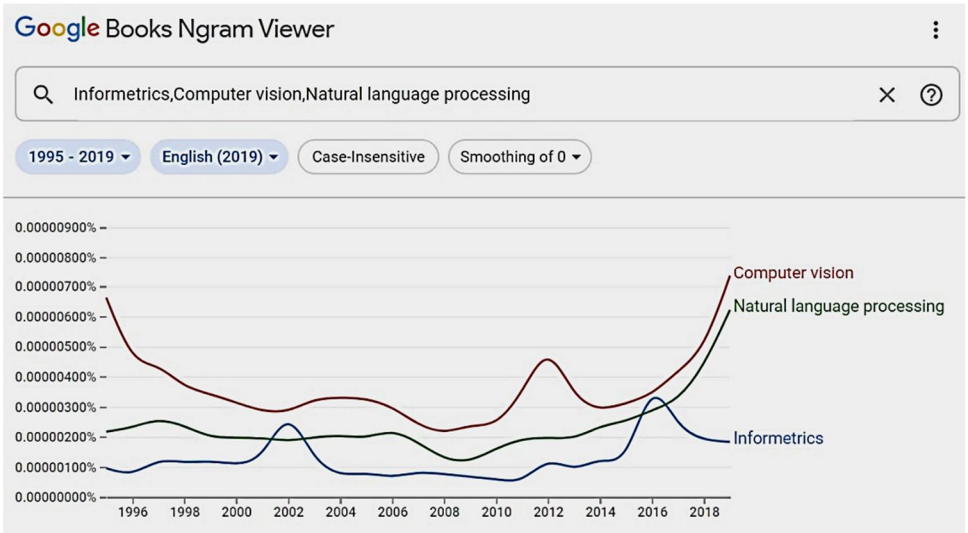

Рис. 9. Динаміка сутностей (кластер Computer Vision, Natural language processing, Informetrics)

Приклад формування та кластеризації кореляційних мереж на основі медичних даних

У даному прикладі розглянуто мережу країн, в яких поширено вірус COVID-19 [7]. Джерелом даних для дослідження виступили дані, зібрані Всесвітньою організацією охорони здоров'я [8, 9]. Було використано набір даних щоденної динаміки захворюваності та смертності (https://ourworldindata.org/coronavirus-source-data) для визначення векторів динаміки пандемічного процесу в різних країнах за певний період (обраний з 15.03.2020 по 31.07.2020). У результаті аналізу було сформовано кореляційну матрицю з 50 вузлами (що відповідають країнам) і виконано кластерний аналіз (рис. 10).

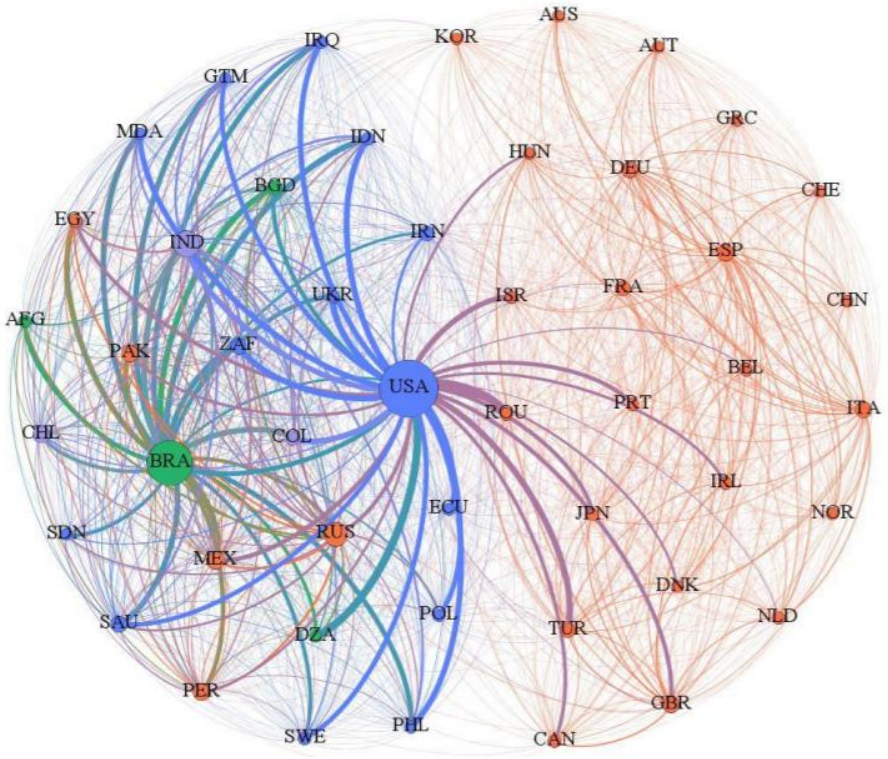

Рис. 10. Направлена зважена кореляційна мережа країн (коди країн відповідно до ISO 5). Колір зв'язків визначається кольором вузлів

Таким чином, отримано групи країн, в яких захворюваність перебігала подібним чином, що дозволяє передбачити наступні хвилі розповсюдження захворюваності в країнах.

Запропонований метод також можна застосувати і для інших об'єктів/сутностей, таких як:

1) політичні лідери, які характеризуються ставленням до різних сфер суспільного життя;

2) споживачі продукції (тут параметри — продавці, джерела продукції) [7];

3) ЗМІ як змістовні сутності, в цьому випадку параметрами можуть бути слова в заголовках статей, які друкуються в цих виданнях.

Висновки

У роботі запропоновано поняття спрямованої кореляційної мережі, яка визначається динамікою появи термінів у публікаціях, описана методика її формування, кластеризації і візуалізації.

Представлений підхід, на відміну від існуючих, має такі переваги:

— інтуїтивно близькі до реальності правила визначення ваг вузлів і зв'язків;

— надійна математична основа кореляційного аналізу;

— застосування невикористовуваних раніше параметрів, часових рядів динаміки публікацій, які відповідають сутностям (тематикам), що дозволяє групувати суті за тенденціями їхнього розвитку в часі;

— об'єктивність — за «чистоту» даних відповідає dataset;

— відносна простота реалізації (використання готових програмних систем, таких як Gephi, мова R і т.д.).

Дієвість методу продемонстровано на прикладах з даними, що отримані в системі Google Books Ngram Viewer, а також медичних даних Всесвітньої організації здоров'я стосовно захворюваності на COVID-19. Наведена методика може базуватися на даних, отриманих, наприклад, від систем контент-моніторингу, використовуватися в аналітичних системах різного призначення з метою узагальнення безлічі сутностей без явно виражених зв'язків між ними.

1. Ping Luo, Kai Shu, Junjie Wu, Li Wan, and Yong Tan. Exploring Correlation Network for Cheating Detection. *ACM Transactions on Intelligent Systems and Technology*. Vol. 11, Issue 1. February 2020. Article No. 12. P. 1–23. https://doi.org/10.1145/3364221

2. Saxena A., Prasad M., Gupta A., Bharill N., Patel O.P., Tiwari A., Lin C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664–681. doi:10.1016/j.neucom. 2017.06.053

3. Foreman J. W. Using Data Science to Transform Information into Insight Data Smart. Wiley, 2013.

4. Lande D.V., Snarskii A.O. Networks determined by the dynamics of thematic information streams. *Data Recording, Storage & Processing*. 2020. Vol. 22, No. 1. P. 56–61. DOI: 10.35681/1560-9189.2020.1.1.207784.

5. Lande D., Strashnoy L. Cross-Correlation of Publications Dynamics and Pandemic Statistics. Available at SSRN: https://ssrn.com/abstract=3625725 or DOI: https://dx.doi.org/10.2139/ssrn.3625725 (June 12, 2020). 9 p.

6. Cherven K. Mastering Gephi Network Visualization. Packt Publishing, 2015.

7. Lande D., Strashnoy L. Directed Correlation Networks, Determined by the Dynamics of COVID-19 Distribution in Various Countries. Available at SSRN: http://ssrn.com/abstract=3674041; DOI: https://dx.doi.org/10.2139/ssrn.3674041 (Posted: 28 Aug 2020). 7 p

8. Coronavirus Source Data. URL: https://ourworldindata.org/coronavirus-source-data

9. Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

Надійшла до редакції 10.06.2021