DOI: 10.35681/1560-9189.2020.1.1.207784

УДК 004.067

Д. В. Ланде[1], А. О. Снарський[2,1]
[1] Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113, Київ, Україна
[2] Національний технічний університет України «КПІ імені Ігоря Сікорського»
Проспект Перемоги, 37, 03056, Київ, Україна

Мережі, що визначаються динамікою тематичних інформаційних потоків

Запропоновано методику формування, кластеризації і візуалізації так званих кореляційних мереж. Зв'язки між вузлами таких мереж відповідають значенням кореляцій між векторами — наборами параметрів, що відповідають цим вузлам. Для побудови мережевих структур для кожного вузла (тематики) формуються вектори — масиви чисел, що відповідають тематичним документальним добіркам. Для цього передбачено застосування системи контент-моніторингу соціальних медіа. Наведений підхід, на відміну від існуючих, має такі переваги як відносно низька розмірність векторів-параметрів, що відповідають тематикам; незалежність від мови документів — вектори параметрів визначаються лише запитами до системи контент-моніторингу, які можуть містити слова, наведені різними мовами; відносна простота реалізації. Наведена методика може застосовуватися в інформаційно-аналітичних системах різного призначення для аналізу масивів сутностей без явно виражених зв'язків між ними. Кореляційні мережі можна розглядати як основу побудови ймовірнісних мереж і застосування технологій нечітких семантичних мереж для подальшого проведення сценарного аналізу.

Ключові слова: *кореляційна мережа, динаміка інформаційних потоків, система контент-моніторингу, візуалізація мережевих структур, кластерний аналіз, модулярність.*

Вступ

Сучасні інформаційні технології неможливо уявити без методів і засобів обробки мережевих структур, але не завжди ці структури виражені явно. Зрозуміло, якщо йдеться щодо явних мереж, вузлів і зв'язків між ними, то проблем не виникає. А ось як побудувати мережу, щоб застосувати великий спектр методів і засо-

бів її обробки, отримати й інтерпретувати результати, якщо в розпорядженні у дослідника є лише деякі сутності — вузли, але не визначені ребра — зв'язки між ними? Якщо кожний об'єкт системи можна представити у вигляді однорідного багатовимірного вектора параметрів, то, можливо застосувати методи класифікації або кластерного аналізу для виявлення груп подібних документів. У багатьох відомих моделях інформаційного пошуку документа або масиву документів, які відповідають певній тематиці, зокрема, ставиться у відповідність вектор ваги слів, що входять до нього [1, 2]. У цьому випадку існує декілька метрік визначення відстані між документами, найбільш відома серед яких — евклідова. 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 $n = |G|$ — кількість елементів у множині параметрів. Кореляція між сутностями s_i та s_j (a_{ij}) може визначатися, наприклад, як кореляція, тобто косинус кута між відповідними векторами $\overline{w^i}$ та $\overline{w^j}$:

$$A_{ij} = \frac{\left(\overline{w^i}, \overline{w^j}\right)}{\left|\overline{w^i}\right|\left|\overline{w^j}\right|} = \frac{\sum_{k=1}^{n} w_k^i w_k^j}{\sqrt{\sum_{k=1}^{n}\left(w_k^i\right)^2}\sqrt{\sum_{k=1}^{n}\left(w_k^j\right)^2}}, \quad (1)$$

де A_{ij} — елементи кореляційної матриці суміжності $A = \left\|A_{ij}\right\|$.

Цей підхід до моделювання має декілька недоліків, зокрема, надвелика розмірність простору слів, практична неможливість застосовувати для масивів документів, представлених різними мовами, складність і неоднозначність визначення ваги окреми слів і словосполучень.

У цій роботі пропонується метод, що також ставить у відповідність тематиці (тематичному масиву документів) вектор (так званий вектор динаміки), що відповідає розподілу документів за часом (датами). Більш конкретно, кожній добі ставитьсяя у відповідність число — кількість документів із тематичного масиву. Розмірність цього вектора відповідає кількості діб, довжині інтервалу часу, протягом якого формувався масив тематичних документів.

Мета

Мета цієї роботи — представити методику формування, кластеризації, ранжування вузлів і візуалізації так званих кореляційних мереж, графових структур, зв'язки між вузлами (сутностями, тематиками) яких відповідають значенням кореляцій між наборами параметрів, що відповідають цим сутностям.

При цьому необхідно зауважити, що кореляція напряму не означає причинонаслідкових зв'язків, тому кореляційні мережі неможна розглядати як каузальні, семантичні мапи. Разом із цим, кореляцію, поряд з іншими критеріями можна розглядати як основу ймовірнісних оцінок. Тобто кореляційні мережі можна розглядати як основу побудови ймовірнісних мереж, як основу застосування технологій нечітких семантичних мереж для подальшого проведення сценарного аналізу.

Для побудови мережевих структур для кожної сутності/тематики формуються вектори, що відповідають цим сутностям. Для цього передбачається застосування системи контент-моніторингу соціальних медіа, таких як InfoStream [3]. Такі системи дозволяють отримувати масиви чисел, що відповідають тематичним документальним добіркам. Для отримання цих масивів до системи можна звертатися через інтерфейс користувача шляхом введення запиту, наведеного інформаційно-пошуковою мовою.

Після формування векторів, що відповідають окремим сутностям, формується кореляційна мережа, яку можна розглядати як засіб зберегти та візуалізувати сутності, що об'єктивно зв'язані між собою. Дійсно, можна сформувати вектори динаміки для різних сутностей, зв'язок між якими не завжди явний.

На рис. 1 наведено фрагмент інтерфейсу користувача отримання динаміки, що відповідає тематичці «кібребезпека». Запит до системи англійською мовою: *Cybersecurity*.

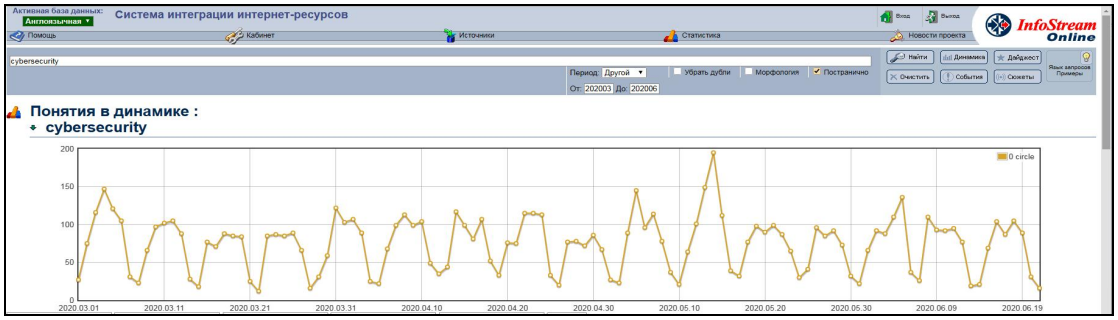

Рис. 1. Фрагмент інтерфейсу системи контент-моніторингу, на якому у вигляді графіка представлений вектор динаміки публікацій за тематикою «кібербезпека»

Метод

Нижче пропонується метод побудови мережі взаємозв'язку сутностей (тематик), що складається з таких етапів.

1. Для кожного поняття формується запит інформаційно-пошуковою мовою системи контент-моніторингу до англомовної частини БД системи. Як приклад розглядається 7 тематик, яким відповідають запити, наведені в таблиці. У системі також визначається період пошуку, що визначає розмірність відповідних векторів динаміки.

Тематики та запити до системи InfoStream

№ сутності	Сутність		Знайдено документів за період 01.04.2020–10.06.2020
1	Кібербезпека	Cyber~security	8979
2	Військові операції	Military~operation	2656
3	Інфляція	Inflation	10649
4	Коронавірус	Coronavirus|COVID-19	844498
5	Вибори в США	Elections&(USA|United~States)	21206
6	Протести в США	Protest&(USA|United~States)	20484
7	Тероризм	Terror	35631

2. У результаті застосування запитів визначається множина векторів динаміки, які відповідають наведеним запитам, аналогічні тим, що наведені на рис. 1. Виконується нормування цих векторів, далі шляхом віконного згладжування елементів (з вікном у 7 діб) забезпечується вилучення тижневої періодичної складової.

3. Обчислюється множина максимальних кросс-кореляцій між отриманими векторами, формується відповідна кореляційна матриця з елементами в позначеннях формули (1):

$$A_{ij}(m) = \max_{m} \frac{\sum_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum_{k=1}^{n-m} \left(w_{k}^{j}\right)^{2}}}. \qquad (2)$$

Функція max використовується з тих міркувань, що близькі за природою процеси можуть мати близьку за динамікою поведінку, але можливо зі зсувом.

4. Здійснюється формування матриці суміжності відповідно до формули (2) і збереження цієї матриці у файлі у форматі CSV. У зв'язку із тим, що в таблиці суміжності існують зв'язки між усіма вузлами, згідно з [4], ігноруються зв'язки, значення яких менші за деякий вибраний поріг. Вибір цього порогу повністю залежіть від досвіду аналітиків. В інформаційній технології, що описується, сформована матриця передається для обробки та візуалізації системі аналізу мережевих структур Gephi (https://gephi.org/) [5]. *Gephi* — це найпоширініша програма візуалізації і аналізу мережевих структур, що забезпечує швидку компоновку, ефективне дослідження даних, а також візуалізацію великомасштабних мереж. Разом із цим, матриця суміжності у форматі CSV для системи Gephi має деякі особливості, які необхідно враховувати (нулі по діагоналі, розташування символів «;» тощо (рис. 2)).

A20

	A	B	C	D	E	F	G	H
1		Cyber_Security	Military_Operation	Inflation	Coronavirus	Elections	Protest_USA	Terror
2	Cyber_Security	0.000	0.975	0.978	0.989	0.984	0.734	0.966
3	Military_Operation	0.975	0.000	0.979	0.980	0.975	0.750	0.965
4	Inflation	0.978	0.979	0.000	0.991	0.964	0.671	0.941
5	Coronavirus	0.989	0.980	0.991	0.000	0.981	0.701	0.950
6	Elections	0.984	0.975	0.964	0.981	0.000	0.763	0.972
7	Protest_USA	0.734	0.750	0.671	0.701	0.763	0.000	0.852
8	Terror	0.966	0.965	0.941	0.950	0.972	0.852	0.000

Рис. 2. Відображення матриці суміжності прикладу в середовищі Excel

5. Здійснюється завантаження цієї матриці у форматі CSV в систему *Gephi*. Ця система має ряд режимів, серед яких для отрмання мережевих характеристик застосовується режим «Лабораторія даних». У цьому режимі, крім звичайних ступенів вузлів матриці, можна розрахувати і їхні значення за PageRank, Hits, модулярності тощо. Крім того існують можливості ранжування вузлів матриці (сутностей) за цими параметрами (рис. 3).

Id	Modularity Class
Cyber_Security	1
Military_Operation	0
Inflation	0
Coronavirus	1
Elections	2
Protest_USA	3
Terror	2

Рис. 3. Фрагмент таблиці в режимі «Лабораторія даних» системи *Gephi*

6. Здійснюється визначення класів модулярності груп об'єктів і подальша кластеризація завантаженої мережевої структури [4, 6]. Модулярність обчислюється як різниця між часткою ребер всередині кластера в розглянутій мережі і очікуваної часткою ребер всередині кластера в мережі, в якій вершини мають ту ж ступінь, що і у вихідній, але ребра розподілені випадково. Модульність мережі можна виразити формулою

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\ \delta\left(c_i, c_j\right), \tag{3}$$

де A_{ij} — елемент матриці суміжності A; m — кількість ребер у графі; k_i, k_j — ступеня вузлів v та w відповідно; δ — дельта Кронекера (показує, чи знаходяться вузли i та j в одному модулі).

7. Візуалізація мережі виконується в системі Gephi. Результати візуалізації мережі сутностей (тематик), що відповідають заданим у таблиці запитам, наведені на рис. 4.

8. На останьому етапі здійснюється експертна інтерпретація результатів.

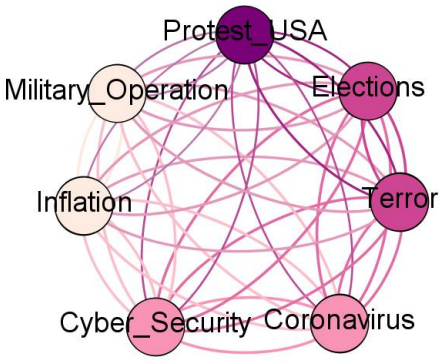

Рис. 4. Мережа сутностей (тематик), у середовищі *Gephi*

Прикладами сутностей, для яких можна застосувати розроблену методику, є наступні.

1. Політичні лідери, які характеризуються відношенням до різних сфер суспільного життя.

2. Споживачі продукції — тут параметри продавці, джерела продукції [4].

3. ЗМІ як змістовні сутності, у цьому разі параметрами можуть бути слова як індикатори «фейків» У заголовках статей, що друкуються у цих виданнях.

Результати візуалізації мережі Telegram-каналів наведено на рис. 4.

Висновки

У роботі описано поняття кореляційної мережі, методику її формування, кластеризації, ранжирування вузлів, яким відповідають вектори динаміки, візуалізації.

Наведений підхід, на відміну від існуючих має такі переваги:

— відносно низька розмірність векторів-параметрів, які відповідають сутностям (тематикам);

— незалежність від мови документів — вектори параметрів визначаються лише запитами до системи контент-моніторингу, які можуть містити слова, наведені різними мовами;

— відносна простота реалізації (можуть застосовуватися готові програмні системи, такі як *Gephi*, *Matlab*, *Excel*, мова *R* тощо).

Наведена методика може застосовуватися в інформаційно-аналітичних системах різного призначення для аналізу масивів сутностей без явно виражених зв'язків між ними.

1. Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. Москва: Вильямс, 2011. 528 с.

2. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. Москва: Либроком (Editorial URSS), 2009. 264 с.

3. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособ. А.Н. Григорьев, Д.В. Ландэ, С.А. Бороденков и др. — Киев: ООО «Старт-98», 2007. 40 с.

4. John W. Foreman. Using Data Science to Transform Information into Insight *Data Smart.* Wiley, 2013.

5. *Ken Cherven.* Mastering Gephi Network Visualization. Packt Publishing, 2015.

6. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навч. посіб. — Київ: ІСЗЗІ КПІ ім. Ігоря Сікрорського, 2018. 300 с.

Надійшла до редакції 30.03.2020