УДК 004(9+056.5)

Кореляційні мережі

Ланде Д.В.

Постановка проблеми

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 засобів її обробки, отримати й інтерпретувати результати, якщо в розпорядженні у дослідника є лише деякі сутності – вузли, але не визначені ребра – зв'язки між ними?

Мета

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

Кореляційну мережу можна розглядати як засіб зберегти і візуалізувати сутності, що об'єктивно зв'язані між собою.

При цьому необхідно зауважити, що кореляція напряму не означає причино-наслідкових зв'язків, тому кореляційні мережі неможна розглядати як каузальні, семантичні мапи. Разом з 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.

Обґрунтування, методологія

Кореляційна мережа формується за таким принципом. Кожному об'єкту (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 в множині параметрів. Кореляція між сутностями s_i і s_j (a_{ij}) визначається як кореляція, тобто косинус кута між відповідними векторами $\overline{w^i}$ та $\overline{w^j}$:

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

Квадратна матриця A з елементами a_{ij}– матриця суміжності кореляційної матриці.

В інформаційній технології, що описується, сформована матриця передається для обробки і візуалізації системі аналізу мережевих структур *Gephi* (https://gephi.org/) [1]. *Gephi*– це найпоширініша програма візуалізації і аналізу мережевих структур, що забезпечує швидку компоновку, ефективне дослідження даних, а також візуалізацію великомасштабних мереж.

Пропонується методика формування і обробки кластерних мереж, що складається із таких етапів:

1. Формування матриці суміжності у відповідності із наведеною формулою і збереження цієї матриці у файлі в форматі CSV.
2. Фільтрація значень, вибір самих «вагомих».
3. Завантаження значень цієї матриці в систему *Gephi.*
4. Ранжування об'єктів [2] в системі *Gephi.*

5. Кластерізація, визначення класів модулярності груп об’єктів [3].
6. Візуалізація мережі в системі *Gephi*.
7. Інтерпретація результатів.

Прикладами сутностей, для яких можна застосувати розроблену методику, такі:

1. Політичні лідери, що характеризуються відношенням до різних сфер суспільного життя (параметри).
2. Споживачі продукції – тут параметри продавці, джерела продукції [4].
3. ЗМІ як змістовні сутності, у цьому разі параметрами можуть бути слова як індикатори «фейків» в заголовках статей, що друкуються у цих виданнях.

Як реалізацію методики було розроблено комплекс програмних модулів, за допомогою якого обробляються новини з мережевих ЗМІ та формується мережа джерел інформації, що допомагає краще її сприймати та обробляти. Як приклад, у доповіді розглядається задача групування каналів месенджера Telegram, що публікують фейкову інформацію. Як параметри цих каналів розглядається G – множина слів, що маркують маніпуляції: $G = \{g_i\}_{i=1}^{|G|}$. Результати візуалізації мережі Telegram-каналів наведено на Рис. 1.

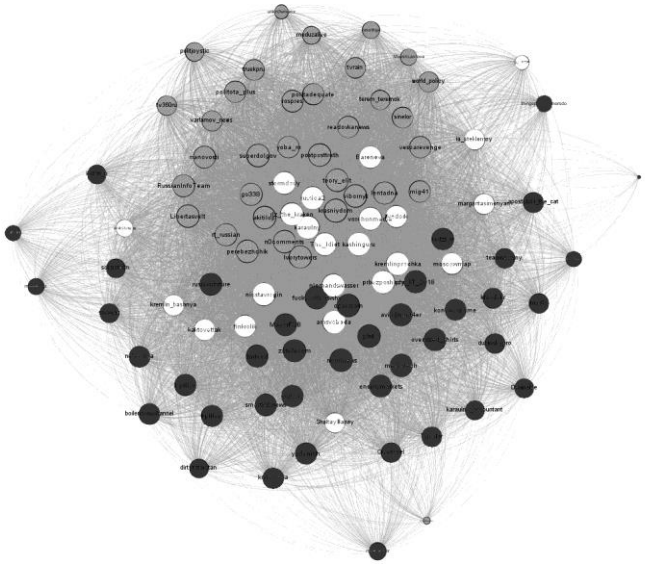

Рисунок 1 – Мережа джерел інформації – каналів месенджера Telegram у середовищі *Gephi*

Висновки

В роботі описано поняття кореляційної мережі, методику її формування, кластеризації, ранжирування вузлів, візуалізації.

Дослідження, проведені на реальному масиві даних дозволили визначити основні три класи модулярності каналів месенджера Telegram, що публікують «фейкові» новини, після чого, аналітики визначили зміст, притаманний цим класам.

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

1. *Ken Cherven.* Mastering Gephi Network Visualization. – Packt Publishing, 2015. ISBN 78-1-78398-734-4.
2. *Снарский А.А., Ландэ Д.В.* Моделирование сложных сетей: учебное пособие. – К.: Инжиниринг, 2015. – 212 с. ISBN 978-966-2344-44-8.
3. *Ланде Д.В., Субач І.Ю., Бояринова Ю.Є.* Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. – К.: ІСЗЗІ КПІ ім. Ігоря Сікрорського, 2018. – 300 с. ISBN 978-966-2577-12-9.
4. *John W. Foreman. Data Smart.* Using Data Science to Transform Information into Insight. – Wiley, 2013. ISBN 111-8-66146-X, 978-1-11866-146-8.