DOI: 10.35681/1560-9189.2019.21.4.199357

УДК 004.067

Д. В. Ланде, О. О. Дмитренко
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Визначення вагових значень зв'язків у мережі термінів

Розглянуто одну із найбільш актуальних проблем комп'ютерного аналізу природної мови — формалізацію та побудову онтологічних моделей предметних областей на основі текстових корпусів заданої тематики. Завдяки глибинному аналізу текстів та обробці природної мови, використовуючи лінгвостатистичні методи та методики обчислювальної лінгвістики, побудовано мережеві моделі предметних областей, для забезпечення кращої взаємодії комунікативних актів, поданих у знаково-словесній формі, та комп'ютерних систем. Зокрема, застосовуючи новий підхід до визначення вагових значень зв'язків у мережі понять, як апробацію було побудовано онтологічну модель для предметної області, що пов'язана з кліматичною надзвичайною ситуацією. Подальший аналіз побудованої моделі дав змогу визначити найбільш впливові та значущі зв'язки між відповідними вузлами у мережі термінів, які відповідають певним поняттям розглянутої предметної області.

Ключові слова: *інформаційних простір, глибинний аналіз тексту, мережева модель, предметна область, мережа термінів, граф горизонтальної видимості, ненаправлена мережа термінів, направлена зважена мережа термінів.*

Вступ

Зазвичай під поняттям «інформаційний простір» розуміють сукупність результатів семантичної діяльності людини у вигляді інформаційних інтернет-ресурсів, де сконцентровано основні результати її комунікативної діяльності. Відомо, що сучасний інформаційний простір характеризується стрімким розвитком динамічних інформаційних масивів і потоків розподілених у веб-просторі — мережі Інтернет.

Та не завжди з масивних інформаційних потоків і величезних об'ємів даних, які вони собою супроводжують, можна виділити необхідну інформацію, яку потребує користувач у відповідь на свій запит. Зокрема, це пов'язано з тим, що такі потоки містять багато зайвих і навіть шумових даних.

Виявилось, і це підтверджується на практиці, що багато задач, які виникають під час роботи з мережевим інформаційним простором, мають багато чого спільного з математичними науками, що відкриває широкі можливості для застосування потужного математичного апарату [1, 2].

Беручи до уваги той факт, що в інформаційних сховищах, розподілених у мережі, накопичуються петабайти текстових даних, то для забезпечення пошуку розміщеної у мережі інформації необхідна розробка нових підходів і методів збирання і опрацювання цих даних. При цьому, безумовно, повинні враховуватися переваги та недоліки вже існуючих моделей і алгоритмів інформаційного пошуку та аналізу [1, 3–5].

Сучасний розвиток технологій дозволяє у деяких випадках знаходити необхідну інформацію в мережах. Але досі залишаються невирішеними проблеми подальшої аналітичної обробки цієї інформації, виокремлення необхідних фактографічних даних, виявлення тенденцій розвитку в окремих предметних областях, взаємозв'язків об'єктів, подій, розпізнавання змістовних аномалій, прогнозування тощо. Більшість із цих проблем — актуальні питання семантичної обробки надвеликих динамічних текстових масивів.

Також задача контент-моніторингу інформації, як адаптація концепції глибинного аналізу текстів і класичних методів контент-аналізу до умов формування та розвитку динамічних інформаційних масивів, зокрема, потоків інформації в мережі Інтернет є актуальною та становить нерозв'язану досі науково-практичну проблему, оскільки процес комп'ютерної обробки даних зазвичай є складним і вимагає високовартісних апаратних і програмних ресурсів та обчислювальних систем, якими не наділені більшість сучасних серверів.

У роботі розглянуто новий підхід до побудови направлених зважених мереж із термінів, що відповідають певним поняттям розглянутої предметної області. Показано, що запропонований підхід дозволяє зважити та виокремити найбільш вагомі зв'язки, які існують між поняттями у відповідному тексті з метою виявлення значущої інформації.

Комп'ютеризована обробка текстового корпусу

Існують різні техніки комп'ютеризованої обробки та аналізу тексту, як форми природної мови. Зважаючи на невпинний та стрімкий розвиток інформаційного простору, виникає потреба у вдосконаленні існуючих підходів і рішень для роботи з ним або розробці нових, більш адаптованих до сучасних тенденцій. У тому числі потребує вдосконалення і сучасне програмне забезпечення, що призначене для обробки і аналізу мережевих інформаційних потоків і масивів. Також варто відзначити, що задача вибору окремих термінів з корпусу текстових документів і автоматизація такого відбору досі залишається відкритою та до кінця невирішеною.

Одним із основних і початкових підготовчих етапів контент-аналізу є відбір, зокрема, тестових джерел і матеріалів для аналізу — інформаційний пошук, ідентифікація та формування корпусу текстових документів, які містять у собі матеріали за наперед заданою тематикою чи темою.

Також важливе значення має етап обробки сформованого текстового корпусу. Авторами використано такі основні послідовні кроки комп'ютеризованого процесу обробки текстових документів як: токенізація, лематизація, вилучення стоп-слів, стемінг і зважування термінів.

Знаково-словесна форма людських комунікативних актів є достатньо складною і, загалом, складається зі слів, які часто мають спільне походження. Мова, що містить різні форми слова, похідні від іншого слова, які використовуються для

вираження різного змісту, називається переплетеною мовою (Inflected Language). Зрозуміло, що такі словоформи мають спільну основу. Тож згадані вище кроки, які були застосовані в процесі обробки текстового корпусу, дають змогу: здійснити попередній лексичний аналіз, розбивши текст на елементарні одиниці (токени, лексеми); привести словоформу до леми — її нормальної (словникової) форми; вилучити стоп-слова, які не мають ніякого смислового навантаження, тобто є інформаційно-неважливими.

Стоп-словник, який використовувався в межах цієї роботи, був сформований на основі різних стоп-словників, які доступні за посиланнями:

— https://code.google.com/archive/p/stop-words/downloads/;

— http://www.textfixer.com/tutorials/common-english-words.php.

Також сформований стоп-словник доповнювався іншими стоп-словами, які були виявлені експертами в межах досліджуваної області.

Для створення програмної реалізації запропонованих і розглянутих підходів і методів використовується мова програмування Python, а також окремі функції спеціалізованої надбудови — модуля NLTK (Natural Language Toolkit open source library).

Після етапів, описаних вище, щоб об’єднати слова, які мають спільний корінь, з метою нормалізації тексту, пропонується здійснити процес стематизації (обробки алгоритмом стемінгу) [6] — скорочення слова до основи шляхом відкидання афіксів (флексій, морфем-афіксів, постфіксів, суфіксів і префіксів), що формують похідні форми слова [7].

Для створення програмної реалізації, було використано стример PorterStemmer (окрема функція спеціалізованої надбудови — модуля NLTK), що розроблений на мові програмування Python та реалізує алгоритм стемінгу Мартіна Портера [8, 9]. Алгоритм Портера є де-факто стандартним алгоритмом стемінгу для англійської мови. Описані вище кроки дають змогу нормалізувати текст корпусу.

Після представлених вище попередніх етапів обробки текстового корпусу здійснюється процес зважування та виокремлення ключових термінів. Як вагові значення термінів, для формування часового ряду як функції, яка ставить у відповідність терміну число, використовується модифікація класичного статистичного вагового показника важливості терміну TF-IDF (з англ. Term Frequency — частота терміну, Inverse Document Frequency — обернена частота документа) [10, 11], а саме — GTF (Global Term Frequency — глобальна частота терміну) [12].

Цей підхід дозволяє інформаційно-важливим в глобальному контексті елементам тексту мати високий статистичний показник важливості.

Визначення зв’язків у мережі термінів

У зв’язку зі складністю природної мови також не менш складною та відкритою проблемою є визначення синтаксичних зв’язків між термінами в тексті, та визначення напрямків таких зв’язків.

У роботі представлено нові підходи до визначення напрямків зв’язків між вузлами ненаправленої мережі, побудованої із термінів тематичного текстового масиву.

Для побудови направленої мережі термінів, як термінологічної онтології певної предметної області, авторами розглянуто та застосовано модифікований підхід

до побудови мереж на основі часового ряду — модифікований алгоритм графа горизонтальної видимості (Directed Horizontal Visibility Graph algorithm — DHVG) [13]. Сам алгоритм графа горизонтальної видимості (Horizontal Visibility Graph algorithm — HVG) [14–16], у свою чергу, є розширенням стандартного алгоритму графа видимості (Visibility Graph algorithm — VG) [17].

Ненаправлена мережа термінів з використанням алгоритму горизонтальної видимості будується в два етапи [18]. Перший етап полягає в тому, що на горизонтальній осі відмічається ряд вузлів, кожен з яких відповідає термінам у тому порядку, в якому вони з'являються в тексті; а по вертикальній осі відкладаються вагові значення — числові оцінки x_i. На другому етапі будується граф горизонтальної видимості. Вважається, що два вузли t_i та t_j, які відповідають елементам часового ряду x_i та x_j, знаходяться у горизонтальній видимості тоді і тільки тоді, коли

$$x_k < \min(x_i, x_j)$$

для всіх t_k, таких, що $t_i < t_k < t_j$.

Отримана ненаправлена мережа термінів буде називатися графом горизонтальної видимості (рис. 1).

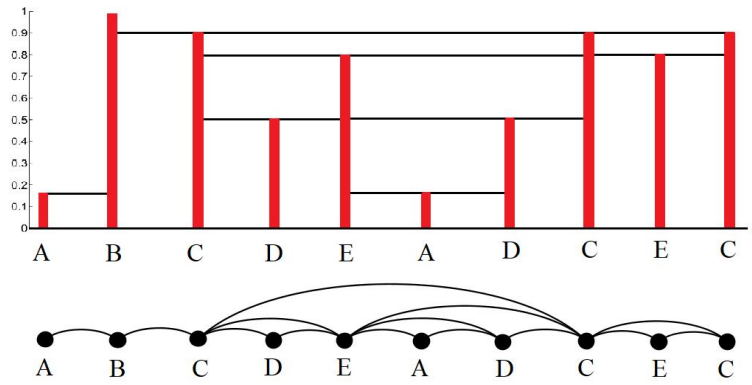

Рис. 1. Етапи побудови компактифікованого графа горизонтальної видимості [18]

Тож розглянутий алгоритм графа горизонтальної видимості дозволяє будувати ненаправлені мережеві структури на основі текстів у випадку, коли окремим словам або словосполученням поставлені у відповідність числові вагові значення.

Для визначення напрямків зв'язків було запропоновано наступний підхід. Нехай G — ненаправлена мережа термінів побудована за принципом, що описаний вище: $G := (V,T)$, де V — множина вузлів; T — множина невпорядкованих пар вузлів з V, які відповідають зв'язкам між вузлами. Вважається, що $\forall_{i,j} : (t_i, t_j) \in T$ — причинно-наслідковий зв'язок — існує в напрямку від вузла t_i до t_j, якщо в реченні термін, якому відповідає вузол t_i, зустрічається раніше ніж термін, якому відповідає вузол t_j. У роботі [13] було досліджено, що описане вище правило, порівняно з іншими, які були запропоновано, більш точно відображає напрямок зв'язків, що існують між термінами в розглянутому тексті. Тобто напрямки зв'язків, визначених за цим правилом, більш точно відображають зміст тексту на думку експертів.

Визначення вагових значень зв′язків у мережі термінів

Також не менш складною та відкритою проблемою є визначення вагових значень зв'язків у мережі термінів.

У цій роботі пропонується новий підхід до визначення вагових значень зв'язків, які встановлені між вузлами направленої мережі, побудованої із термінів тематичного текстового масиву за допомогою алгоритму, що був описаний вище.

Загальний принцип на рівні графа описується наступним чином: вершини графа, що відповідають однаковим термінам побудованої на попередньому етапі направленій мережі, об'єднуються («зшиваються», «склеюються»). Оскільки будь-який граф визначається матрицею суміжності, то задача визначення вагових значень зв'язків зводиться до конкатенації стовпців і відповідних рядків — зваженої компактифікації графа горизонтальної видимості [18].

Більш формально процес визначення вагових значень зв'язків у мережі термінів виглядає наступним чином. Нехай D — направлена мережа термінів, побудована за принципом, що описаний вище: D:= (V,E), де V — множина вузлів, E — множина впорядкованих пар вузлів з V, які відповідають причинно-наслідковим зв'язкам між вузлами. Нехай квадратна матриця A розміру n, в якій значення елемента a_{ij} рівне 0 або 1 залежно від того, чи існує ребро (дуга) в напрямку від вершини i до вершини j. Нехай $T = \{t_1,\ldots,\ t_m\}$ — множина вузлів, що відповідають однаковим термінам у тексті ($1 \le m \le n$). Очевидно, що кожному вузлу t_k ($1 \le k \le m$) із множини T відповідає стовпець a_{ik} та рядок a_{kj} матриці A. Тож відповідні елементи стовпців (рядків) усіх спільних вузлів підсумовуються та записуються в новий стовпець (рядок) — w_{ik} (w_{kj}, відповідно), формуючи нову матрицю W. В отриманій у результаті вищеописаного процесу конкатенації матриці W значення елемента w_{ij} дорівнює числу ребер з i-ї вершини графа до j-ї вершини.

Також для вищеописаного процесу конкатенації стовпців (рядків) використовується так звана хеш-таблиця подібних або ж синонімічних термінів, які відповідають однаковим поняттям, сформована експертами в тій предметній області, яка розглядається. Така таблиця дозволяє додатково об'єднати вузли, що відповідають однаковим за змістом термінам у тексті.

Отримана матриця W визначає орієнтований зважений граф, сформований з вершин, що відповідають унікальним термінам у розглянутому тексті. Вагове значення ребра, що з'єднує вершину i з вершиною j визначається кількістю появ терміна t_i перед терміном t_j в тексті (кількістю появ елемента часового ряду t_i перед елементом t_j).

Виклад основного матеріалу

Запропонований підхід для визначення напрямків і вагових значень зв'язків у направлених мережах термінів був апробований на прикладі корпусу документів, тематично пов'язаного з кліматичною надзвичайною ситуацією (climate emergency).

Для проведення дослідження обраної предметної області було використано вільну доступну пошукову систему, яка індексує повний текст наукових публіка-

цій — Google Scholar (https://scholar.google.com). На цьому етапі було вивантажено анотації перших 100 статей за запитом «climate emergency».

Відповідно до вищеописаних етапів було здійснено обробку обраного текстового корпусу та виокремлено ключові терміни (табл. 1).

Таблиця 1. Топ-21 ключових (найбільш вагомих) термінів для корпусу «climate emergency»

№№	Термін	Вагове значення	№№	Термін	Вагове значення
1	climat	0,0628	12	challeng	0,0038
2	emerg	0,058	13	effect	0,0038
3	chang	0,017	14	nation	0,0038
4	global	0,0082	15	intern	0,0038
5	action	0,0068	16	human	0,0038
6	respons	0,0058	17	safeti	0,0038
7	health	0,0051	18	studi	0,0038
8	polit	0,0048	19	manag	0,0034
9	univers	0,0044	20	polici	0,0034
10	declar	0,0044	21	time	0,0034
11	geoengin	0,0038			

Побудувавши направлену зважену мережу термінів із використанням запропонованого підходу, було отримано результати, які наведено на рис. 2. Візуалізація побудованих мереж термінів здійснювалася за допомогою засобів програмного забезпечення для моделювання та візуалізації графів — Gephi (https://gephi.org).

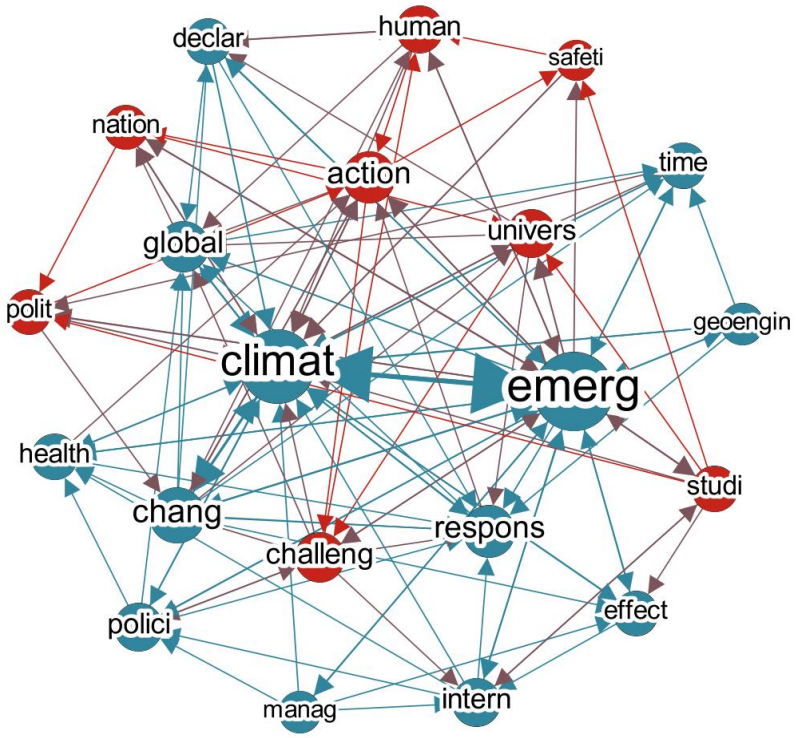

Рис. 2. Направлена зважена мережа термінів, яка побудована для предметної області «climate emergency»

Також за допомогою засобів програмного забезпечення Gephi було отримано такі параметри побудованої мережі: загальна кількість вузлів мережі — 21; зв'язків — 129; середній коефіцієнт кластеризації — 0,525; середня довжина шляху — 1,738; щільність мережі — 0,307; кількість зв'язаних компонент — 1; середня степінь — 6,143.

У табл. 2 представлено список найбільш впливових та значущих зв'язків між відповідними вузлами у мережі термінів, що відповідають певним поняттям розглянутої предметної області.

Таблиця 2. Топ-27 значущих зв'язків для корпусу «climate emergency»

№№	Вихідний вузол	Цільовий вузол	Вагове значення
1	climat	emerg	89
2	emerg	climat	80
3	climat	chang	33
4	chang	emerg	16
5	chang	climat	15
6	global	climat	15
7	safeti	climat	9
8	action	climat	8
9	declar	climat	8
10	emerg	univers	8
11	emerg	global	8
12	emerg	studi	8
13	respons	climat	8
14	emerg	action	7
15	emerg	manag	6
16	emerg	declar	6
17	emerg	polit	6
18	polit	climat	6
19	effect	climat	5
20	emerg	geoengin	5
21	emerg	nation	5
22	emerg	intern	5
23	emerg	respons	5
24	geoengin	climat	5
25	health	emerg	5
26	intern	emerg	5
27	nation	climat	5

Провівши аналіз отриманих результатів, було встановлено, що найбільш впливовими та значущими зв'язками у мережі термінів побудованої для предметної області, пов'язаної із кліматичною надзвичайною ситуацією, є: «climat → emerg», «emerg → climat», «climat → chang», «chang → emerg», «chang → climat» та «global → climat».

Висновки

Розглянуто та застосовано основні підходи до комп'ютеризованої обробки та аналізу текстових документів.

Застосовуючи новий підхід до визначення напрямків і вагових значень зв'язків у мережі термінів, запропонований у цій роботі, було побудовано онтологічну модель для предметної області, пов'язаної з кліматичною надзвичайною ситуацією (climate emergency).

Направлені зважені мережі термінів, побудовані за допомогою запропонованого підходу, можна використовувати як основу для автоматизованої побудови онтологічних моделей предметних областей. Також результати роботи можуть бути використані під час створення персональних пошукових інтерфейсів для користувачів інформаційно-пошукових систем, а також у системах навігації у базах даних. Це повинно допомогти користувачам таких систем спростити процес пошуку релевантної інформації.

Оскільки задача підвищення точності визначення напрямків зв'язків і їхніх вагових значень у ненаправлених мережах термінів залишається актуальною, планується продовжити роботу в цьому напрямку, розвиваючи нові та модифікуючи існуючі підходи.

Для аналізу текстів застосовано виключно статистичні методи, проте надалі авторами планується представити результати аналізу із застосуванням більш широкої обробки природної мови, такої як виокремлення частин мови (Part-of-speech tagging), синтаксичний аналіз та інші типи лінгвістичного аналізу.

1. Снарский А.А., Ландэ Д.В. Моделирование сложных сетей: учеб. пособ. — Киев: ООО «Инжиниринг», 2015. 212 с. ISBN 978-966-2344-44-8

2. Додонов А.Г., Ландэ Д.В., Путятин В.Г. Компьютерные сети и аналитические исследования. Киев: ИПРИ НАН Украины, 2014. 486 с. ISBN 978-966-02-7422-8.

3. Ланде Д.В., Субач І.Ю., Боярінова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навч. посіб. — Київ: ІСЗЗІ «КПІ ім. Ігоря Сікорського», 2018. 297 с.

4. Frakes W.B., Baeza-Yates R. Information retrieval data structures and algorithms. Prentice Hall, Englewood Cliffs, New Jersey, 1992. 512 p.

5. Manning C.D., Raghavan P., & Schütze H. An Introduction to Information Retrieval. Cambridge University Press, 2009. P. 22–36.

6. Lovins J.B. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics.* 1968. **11**(1–2). P. 22–31.

7. Jongejan B., & Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, P. 145–153. Association for Computational Linguistics, Singapore (2009).

8. Porter M.F. An algorithm for suffix stripping. *Program.* 1980. **14**(3). P. 130–137. doi: 10.1108/eb046814

9. Willett P. The Porter stemming algorithm: then and now. *Program*. 2006. **40**(3). P. 219–223. doi: 10.1108/00330330610681295.

10. Salton G., & Buckley C. Term-weighting approaches in automatic text retrieval. *Information processing & management.* 1988. **24**(5). P. 513–523. doi:10.1016/0306-4573(88)90021-0.

11. Rajaraman A., & Ullman J.D. Mining of massive datasets. Cambridge University Press, 2011.

12. Lande D.V., Dmytrenko O.O., Snarskii A.A. Transformation texts into complex network with applying visability graphs algorithms. In: CEUR Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn:de:0074-2318-4. Selected Papers of the XVIIIШ International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). 2018. Vol. 2318. P. 95–106.

13. Ланде Д.В., Дмитренко О.О., Радзієвська О.Г. Визначення напрямків зв'язків у мережі термінів. Інформаційні технології та безпека. Матеріали XIX Міжнародної науково-практичної конференції «ІТБ-2019». Київ: ООО «Инжиниринг», 2019. С. 103–112.

14. Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs: Exact results for random time series. *Physical Review E*. 2009. **80**(4). doi: 10.1103/PhysRevE.80.046103.

15. Gutin G., Mansour T., & Severini S. A characterization of horizontal visibility graphs and combinatorics on words. *Physica A: Statistical Mechanics and its Applications*. 2011. **390**(12). P. 2421–2428. doi: 10.1016/j.physa.2011.02.031.

16. Bezsudnov I.V., Snarskii A.A. From the time series to the complex networks: The parametric natural visibility graph. *Physica A: Statistical Mechanics and its Applications*. 2014. **414**. P. 53–60. doi: 10.1016/j.physa.2014.07.002.

17. Lacasa L., Luque B., Ballesteros F., Luque J., Nuno J.C. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*. 2008. **105**(13). P. 4972–4975. doi: 10.1073/pnas.0709247105

18. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text. In: 2013 12th Mexican International Conference on Artificial Intelligence. 2013. P. 209–215.

Надійшла до редакції 06.12.2019