УДК 004.7

МЕТОД ПОБУДОВИ НАПРАВЛЕНИХ ЗВАЖЕНИХ МЕРЕЖ ТЕРМІНІВ НА ОСНОВІ ТЕКСТОВИХ КОРПУСІВ

Д. В. Ланде[1,2, а], О. О. Дмитренко[1, б]

[1] *Інститут проблем реєстрації інформації НАН України*

[2] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»*

Анотація

В цій роботі розглядається одна з найбільш актуальних проблем комп'ютерного аналізу природньої мови – формування термінологічних онтологій предметних областей на основі текстових корпусів заданої тематики. Мережа термінів, вузли якої відповідають окремим словам та словосполученням у тексті, а ребра – зв'язкам між ними, розглядатися в якості онтологічної моделі. Для побудови направленої мережі із термінів, що відповідають певним поняттям обраної предметної області, в цій роботі застосовується новий підхід до побудови мереж на основі часового ряду – модифікований алгоритм графа горизонтальної видимості (Directed Horizontal Visibility Graph algorithm – DHVG). Також запропоновано новий метод розрахунку вагових значень зв'язків у направленій мережі термінів. Апробацію запропонованого методу було здійснено на основі текстового корпусу з інформаційних повідомлень за темою «COVID-19». Також було розглянуто корпус тієї ж теми, сформований з анотацій наукових статей, що були вивантажені з вільної доступної пошукової системи Google Scholar. Результатом роботи стало візуальне представлення мереж термінів для текстових корпусів за темою «COVID-19».

Ключові слова: предметна область, термінологічна онтологія, граф горизонтальної видимості, направлена зважена мережа термінів, COVID-19

Вступ

Стрімкий розвиток інформаційно-комунікаційних технологій та глобалізація інформаційного простору призвели до не менш стрімкого розвитку інформаційних ресурсів у мережі Інтернет. При збиранні й аналізі надвеликих обсягів неструктурованих даних, що вони містять, виникають проблеми, пов'язані з обробкою, пошуком і навігацією. Тож у зв'язку зі збільшенням динамічних інформаційних потоків, що розподілені в веб-мережі, формалізоване представлення знань та побудова термінологічних онтологій предметних областей на основі масиву текстових документів заданої тематики є актуальними [1]. Одним із видів такої онтологічної моделі може бути мережа, що побудована із ключових термінів [2, 3]. В ній вузли відповідають окремим поняттям, а ребра – зв'язкам між ними. Під час побудови таких мереж відкритою та до кінця не вирішеною проблемою є визначення та виокремлення базових об'єктів (ключових термінів – слів та словосполучень), встановлення напрямків зв'язків між термінами та визначення вагових значень цих зв'язків. Не менш важливою є автоматизація вищезгаданих процесів.

В цій роботі розглядається та застосовується новий метод побудови направлених зважених мереж термінів на основі текстових корпусів [4].

[а] dwlande.o@gmail.com

[б] dmytrenko.o@gmail.com

1. Комп'ютеризована обробка текстового корпусу та виокремлення ключових термінів

Початковим етапом формування мережі термінів, пов'язаної з певною предметною областю, є формування корпусів текстових документів. На цьому етапі із вільнодоступних пошукових систем, відповідно до пошукового запиту, вивантажуються тестові документи або їх короткі анотації за обраною тематикою.

Наступний етап комп'ютеризованої обробки текстового корпусу включає попередній лексичний аналіз – розбиття тексту на елементарні одиниці (токени або лексеми), лематизацію [5, 6], вилучення стопслів, які є інформаційно-неважливими та не мають ніякого смислового навантаження, та стематизацію [7, 8].

Наступним кроком є зважування й виокремлення ключових термінів. В якості функції, яка ставить у відповідність слову число, в цій роботі використовується статистичний показник важливості терміна – глобальний TF (Global Term Frequency – GTF) [9], що визначається відношенням загальної кількості появи терма у всіх документах корпусу до загальної кількості термінів у документах корпусу. Використання цього показника дає змогу уникнути ситуації, що виникає під час роботи з текстовим корпусом заздалегідь визначеної тематики, коли інформаційно-важливий терм зустрічається майже у кожному документі корпусу, але має при цьому низький ваговий показник TF (Term Frequency) [10, 11]. В той час,

Рис. 1. Приклад побудови DHVG

GTF навпаки – дозволяє інформаційно-важливим в глобальному контексті елементам тексту мати високий статистичний показник важливості.

Присвоївши кожному слову число у порядку появи їх у тексті, отримуємо часовий ряд, який далі буде трансформований у мережу зі слів.

2. Побудова направленої мережі зі слів

У цій роботі для побудови направленої мережі зі слів розглядається новий підхід до визначення напрямків зв'язків між вузлами ненаправленої мережі, побудованої із термінів текстового корпусу – модифікований алгоритм графа горизонтальної видимості – направлений граф горизонтальної видимості (Directed Horizontal Visibility Graph algorithm – DHVG) [12].

Щоб побудувати ненаправлену мережу для послідовності слів та їх вагових значень, в даній роботі використовується алгоритм побудови графа горизонтальної видимості (Horizontal Visibility Graph algorithm – HVG) [13, 14, 15]. Алгоритм графа горизонтальної видимості є розширенням стандартного алгоритму графа видимості [16]. Цей алгоритм ставить у відповідність часовому ряду граф, сформований з його елементів (слів).

Для визначення напрямків зв'язків у ненаправленій мережі зі слів був запропонований наступний підхід [12]. У графі горизонтальної видимості ненаправлений зв'язок між вузлами, якщо він існує, будується у напрямку від t_i до t_j, якщо у реченні (сформованому часовому ряді) термін, якому відповідає вузол t_i зустрічається раніше ніж термін, якому відповідає вузол t_j. У результаті, отримуємо направлений граф горизонтальної видимості (DHVG) як направлену мережу зі слів (рис. 1).

3. Метод розрахунку вагових значень зв'язків у направленій мережі термінів

В цій роботі пропонується новий метод визначення вагових значень зв'язків у направленій мережі термінів побудованої за допомогою вищеописаного підходу.

На рівні теорії графів загальний принцип полягає в наступному: вершини графа, що відповідають однаковим термінам побудованої на попередньому етапі направленій мережі об'єднуються ("зшиваються", "склеюються") (рис. 2). Оскільки будь-який

Рис. 2. Приклад визначення вагових значень зв'язків у направленій мережі

граф визначається матрицею суміжності, то задача визначення вагових значень зв'язків зводиться до конкатенації стовпців та відповідних рядків – зваженої компактифікації графа горизонтальної видимості [15].

Іншими словами, 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 A – квадратна матриця розміру n, в якій значення елемента a_{ij} рівне 0 або 1 в залежності від того, чи існує ребро (дуга) в напрямку від вершини i до вершини j. Нехай $T = \{t_i,..,t_m\}$ – множина вузлів, що відповідають однаковим термінам у тексті ($1 \leq \mathrm{m} \leq \mathrm{n}$). Очевидно, що кожному вузлу t_k ($1 \leq \mathrm{k} \leq \mathrm{m}$) із множини T відповідає стовпець a_{ik} та рядок a_{kj} матриці A. Тож відповідні елементи стовпців (рядків) всіх спільних вузлів підсумовуються й записуються у новий стовпець (рядок) – w_{ik} (w_{kj}, відповідно), формуючи нову матрицю W. В отриманій урезультаті вищеописаного процесу конкатенації матриці W значення елемента w_{ij} рівне числу ребер з i-ї вершини графа в j-у вершину й визначається кількістю зустрічань терміна t_i перед терміном t_j у тексті.

4. Візуалізація й аналіз результатів

Запропонований підхід для визначення напрямків та вагових значень зв'язків у ненаправлених мережах термінів був апробований на прикладі корпусів заздалегідь вибраних документів, тематично пов'язаних із COVID-19. Із найбільшої пошукової системи – Google Search (google.com) було вивантажено 440 реферативних інформаційних повідомлень за темою «COVID-19». Ще один корпус за тією ж темою було сформовано з використанням вільної доступної пошукової системи, яка індексує повний текст наукових публікацій – Google Scholar (scholar.google.com), звідки було вивантажено 430 анотацій статей.

Відповідно до вищеописаних етапів було здійснено обробку обраного текстових корпусів й виокремлено ключові терміни (Таблиця 1).

Побудувавши направлені зважені мережі термінів з використанням запропонованого підходу було отримано наступні результати (рис. 3, 4).

Табл. 1. Топ-26 ключових (найбільш вагомих) термінів для корпусів сформованих на основі Google Search та Google Scholar

	Google Search		Google Scholar	
№	Слово	GTF	Слово	GTF
1	covid-19	0.1783	covid-19	0.1118
2	diseas	0.0228	diseas	0.0262
3	health	0.02	china	0.0229
4	case	0.014	case	0.0159
5	inform	0.0092	outbreak	0.0145
6	outbreak	0.008	patient	0.014
7	confirm	0.0061	infect	0.0121
8	spread	0.0059	wuhan	0.0115
9	public	0.0059	sever	0.0102
10	symptom	0.0059	sars-cov-2	0.0098
11	respons	0.0058	respiratori	0.0095
12	global	0.0051	health	0.0091
13	infect	0.0051	spread	0.0089
14	respiratori	0.0051	pneumonia	0.0084
15	resourc	0.0049	report	0.0081
16	test	0.0049	acut	0.0078
17	caus	0.0048	emerg	0.0074
18	countri	0.0048	clinic	0.0069
19	depart	0.0045	syndrom	0.0067
20	data	0.0044	caus	0.0063
21	peopl	0.004	epidem	0.0061
22	offici	0.0041	ct	0.0054
23	death	0.0039	confirm	0.0053
24	patient	0.0039	pandem	0.0053
25	emerg	0.0039	transmiss	0.005
26	report	0.0036	background	0.0045

Рис. 3. Направлена зважена мережа термінів побудована для корпусу текстових документів із Google Search

Рис. 4. Направлена зважена мережа термінів побудована для корпусу текстових документів із Google Scholar

Також за допомогою засобів програмного забезпечення Gephi (gephi.org) [17] були отримані такі параметри побудованих мереж. Для мережі побудованої на основі корпусу отриманого із Google Search загальна кількість вузлів мережі – 26; зв'язків – 234; середній коефіцієнт кластеризації – 0.501; середня довжина шляху – 1.64; щільність мережі – 0.36; кількість зв'язаних компонент – 1; середня степінь – 9. Для мережі побудованої на основі корпусу отриманого із Google Scholar загальна кількість вузлів мережі – 26; зв'язків – 441; середній коефіцієнт кластеризації – 0.743; середня довжина шляху – 1.322; щільність мережі – 0.678; кількість зв'язаних компонент – 1; середня степінь – 16.96.

У таблиці 2 представлений список значущих зв'язків між відповідними вузлами у мережі термінів, що побудована для корпусу текстових документів із Google Search та Google Scholar.

Висновки

В цій роботі було розглянуто та застосовано основні підходи до комп'ютеризованої обробки та аналізу текстових корпусів. Також розглянуто та застосовано новий підхід до побудови мереж на основі часового ряду – модифікований алгоритм графа горизонтальної видимості (Directed Horizontal Visibility Graph algorithm – DHVG). Запропоновано новий метод розрахунку вагових значень зв'язків у направленій мережі термінів. Апробацію розглянутого методу було здійснено на основі текстового корпусу за темою «COVID-19».

Мережа, що побудована повністю автоматично із ключових термінів на основі масиву текстових документів заданої тематики, може розглядатися як основа для подальшого автоматизованого формування термінологічних онтологій за участю експертів.

Табл. 2. Топ-26 ключових (найбільш вагомих) термінів для корпусів сформованих на основі Google Search та Google Scholar

	Google Search			Google Scholar		
№	Вихідний вузол	Цільовий вузол	Weight	Вихідний вузол	Цільовий вузол	Weight
1	diseas	covid-19	134	diseas	covid-19	270
2	covid-19	diseas	101	covid-19	diseas	251
3	health	covid-19	98	covid-19	china	213
4	covid-19	health	70	wuhan	china	119
5	case	covid-19	59	patient	covid-19	117
6	covid-19	case	52	china	covid-19	113
7	inform	covid-19	39	case	covid-19	106

Використання запропонованого методу ддя побудови направлених зважених мереж із термінів дасть змогу виокремити значущу інформацію із масиву неструктурованих даних та спростити процес інформаційного пошуку та автоматичного реферування. Також запропонований метод може бути використаний під час розробки інформаційно-аналітичних систем та створення персональних пошукових інтерфейсів для користувачів інформаційно-пошукових систем, а також у системах навігації у базах даних.

Перелік використаних джерел

1. Navigli R., Velardi P., Gangemi A. Ontology learning and its application to automated terminology translation // IEEE Intelligent systems. — 2003. — Vol. 18, no. 1. — P. 22–31.
2. Faber P., León P., Prieto J. A. Semantic relations, dynamicity, and terminological knowledge bases // Current Issues in Language Studies. — 2009. — Vol. 1, no. 1. — P. 1–23.
3. Ланде Д.В., Снарский А.А. Подход к созданию терминологических онтологий // Онтология проектирования. — 2014. — Т. 2, № 12. — С. 83–91.
4. Ланде Д.В., Дмитренко О.О. Визначення вагових значень зв'язків у мережі термінів // Реєстрація, зберігання і обробка даних. — 2019. — Т. 21, № 4. — С. 40–48.
5. Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval // Cambridge University Press. — 2009. — P. 22–36.
6. Jongejan B., Dalianis H. Automatic training of lemmatization rules that handle morpho-logical changes in pre-, in-and suffixes alike // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Lan-guage Processing of the Asian Federation of Natural Language Processing. — 2009. — P. 145–153.
7. Porter M. F. An algorithm for suffix stripping // Program. — 1980. — Vol. 14, no. 3. — P. 130–137.
8. Willett P. The Porter stemming algorithm: then and now // Program. — 2006. — Vol. 40, no. 3. — P. 219–223.
9. Lande D.V., Dmytrenko O.O., A.A. Snarskii. Transformation texts into complex network with applying visability graphs algorithms // CEUR Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn:de:0074-2318-4. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). — 2018. — Vol. 2318. — P. 95–106. — online; accessed: `http://ceur-ws.org/Vol-2318/paper8.pdf`.
10. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. Information processing management // Program. — 1988. — Vol. 24, no. 5. — P. 513–523.
11. Rajaraman A., Ullman J. D. Mining of massive datasets. — Cambridge University Press, 2011.
12. Lande D., Dmytrenko O., Radziievska O. Determining the Directions of Links in Undirected Networks of Terms // CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference Information Technologies and Security (ITS 2019). — 2019. — Vol. 2577. — P. 132–145. — Access mode: `http://ceur-ws.org/Vol-2577/paper11.pdf`.
13. Luque B., Lacasa L., Ballesteros F.and Luque J. Horizontal visibility graphs: Exact results for random time series // Physical Review E. — 2009. — Vol. 80, no. 4. — P. 40–48.
14. Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatorics on words // Physica A: Statistical Mechanics and its Applications. — 2011. — Vol. 390, no. 12. — P. 2421–2428.
15. The use of horizontal visibility graphs to identify the words that define the informational structure of a text / D. V. Lande, A. A. Snarskii, E. V. Yagunova, E. V. Pronoza // 2013 12th Mexican International Conference on Artificial Intelligence. — 2013. — P. 209–215.
16. From time series to complex networks: the visibility graph / L. Lacasa, B. Luque, J. Ballesteros, F.and Luque, J.C. Nuño // Proceedings of the National Academy of Sciences. — 2008. — Vol. 105, no. 13. — P. 4972–4975.
17. Cherven K. Mining of massive datasetsNetwork graph analysis and visualization with Gephi. — Packt Publishing Ltd, 2013.