

Метод Побудови Направлених Зважених Мереж Термінів На Основі Текстових Корпусів

^{1,2} Ланде Д.В., ¹ Дмитренко О.О.

¹ Інститут проблем реєстрації інформації НАН України, Київ, Україна

² Національний технічний університет України «Київський політехнічний
інститут імені Ігоря Сікорського», Київ, Україна

12-13 травня 2020 рік

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.

Побудова термінологічної онтології

Одним із видів такої онтологічної моделі може бути мережа, що побудована із ключових термінів:

- ▶ **Вузли** відповідають окремим поняттям (термінам у тексті);
- ▶ **Ребра** – зв'язкам між поняттями

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
- ▶ визначення **вагових значень** цих зв'язків.

Не менш важливою є **автоматизація** вищезгаданих процесів.

Формування та комп'ютеризована обробка текстового корпусу

Початковим етапом формування мережі термінів, пов'язаної з певною предметною областю, є **формування корпусів текстових документів**.

Далі здійснюються такі основні етапи комп'ютеризованого процесу обробки текстових документів:

- ▶ **Токенізація** (попередній лексичний аналіз – розбиття тексту на елементарні одиниці (токени або лексеми));
- ▶ **Лематизація** (приведення словоформи до леми – її нормальної (словникової) форми);
- ▶ **Вилучення стоп-слів** (слова, які не мають ніякого смислового навантаження, тобто є інформаційно-неважливими);
- ▶ **Стематизація** (скорочення слова до основи шляхом відкидання афіксів, що формують похідні форми слова).

Виокремлення ключових термінів

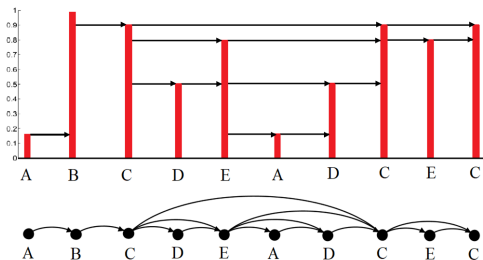
В якості функції, яка ставить у відповідність слову число, в цій роботі використовується статистичний показник важливості терміна – глобальний **TF (Global Term Frequency – GTF)**, що визначається відношенням загальної кількості появи терма у всіх документах корпусу до загальної кількості термінів у документах корпусу.

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

Directed Horizontal Visibility Graph algorithm – DHVG

Щоб побудувати ненаправлену мережу для послідовності слів та їх вагових значень, в даній роботі використовується Horizontal **Visibility Graph algorithm** – HVG.

Для визначення напрямків зв'язків у ненаправленій мережі зі слів та побудови **DHVG** був запропонований наступний підхід. У HVG зв'язок між вузлами, якщо він існує, будується у напрямку від t_i до t_j , якщо у реченні термін, якому відповідає вузол t_i зустрічається раніше ніж термін, якому відповідає вузол t_j .



Метод розрахунку вагових значень зв'язків у DHVG

У 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”), а відповідні числові значення стовпців та рядків графа підсумовуються.

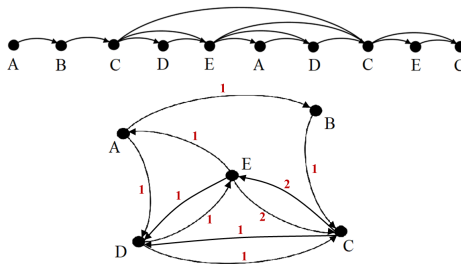


Рис. 2: Визначення вагових значень зв'язків у направленій мережі

Формування корпусів за темою «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.

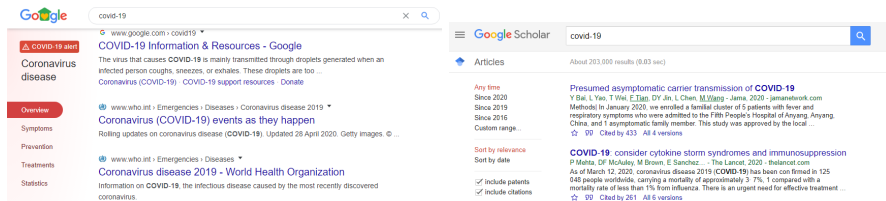


Рис. 3: Пошукові запити у системах Google Search та Google Scholar

Ключові слова

№	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

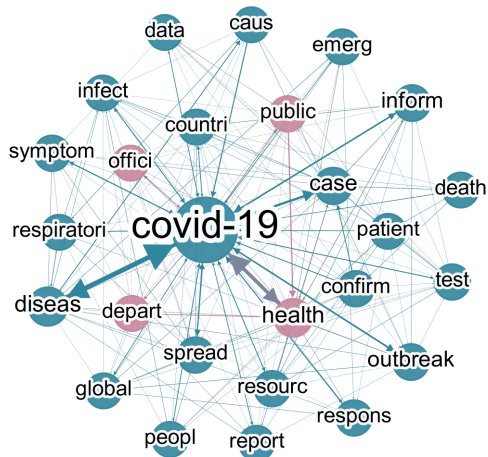
Табл. 1: Топ-16 ключових (найбільш вагомих) термінів для корпусів сформованих на основі Google Search та Google Scholar.

Ключові зв'язки

№	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

Табл. 2: Топ-7 найбільш вагомих зв'язків для корпусів із Google Search та Google Scholar.

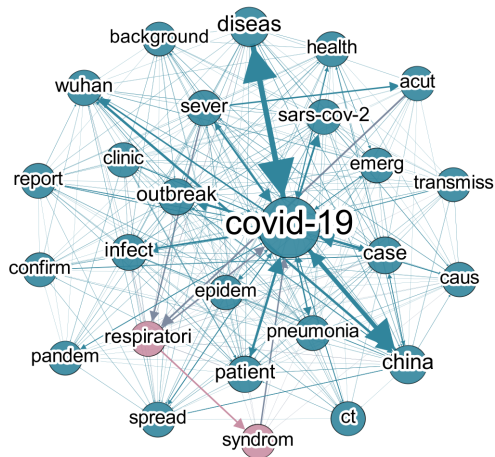
Візуалізація й аналіз результатів (Google Search)



- ▶ Загальна кількість вузлів мережі – 26;
- ▶ Загальна кількість 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.

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

Візуалізація й аналіз результатів (Google Scholar)



- ▶ Загальна кількість вузлів мережі – 26;
- ▶ Загальна кількість 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.

Рис. 5: Направлена зважена мережа термінів побудована для корпусу текстових документів із Google Scholar

Висновки

- ▶ Розглянуто та застосовано основні підходи до комп'ютеризованої обробки та аналізу текстових корпусів.
- ▶ Розглянуто та застосовано новий підхід до побудови направлених мереж на основі часового ряду.
- ▶ Запропоновано новий метод розрахунку вагових значень зв'язків у направлених мережі термінів.
- ▶ Апробацію розглянутого методу було здійснено на основі текстового корпусу за темою «COVID-19».