INFORMATION TECHNOLOGY

DOI: 10.20535/2411-1031.2018.6.2.153486
УДК 004.67

ДМИТРО ЛАНДЕ,
ОЛЕГ ДМИТРЕНКО

СТВОРЕННЯ МЕРЕЖ СЛІВ НА ОСНОВІ ТЕКСТІВ З ВИКОРИСТАННЯМ АЛГОРИТМІВ ГРАФІВ ВИДИМОСТІ

Пропонується метод створення мереж із текстів, так званих мереж слів (Language Network). Із масиву заздалегідь вибраних текстових документів, які описують певну предметну область, виділяються окремі слова та ключові поняття. Використовуючи статистичний показник TF-IDF окремим словам ставляться у відповідність числові вагові значення, і як результат, формується часовий ряд. Використовуючи алгоритми побудови графів видимості як інструмент для аналізу часових рядів, між отриманими ключовими поняттями будується граф предметної області. Для прикладу, в роботі розглядаються актуальні предметні області: “Космічний простір” та “Комп’ютерна графіка”. Для масиву заздалегідь вибраних текстових документів, тематично пов’язаних з поняттям космічного простору та комп’ютерної графіки, застосовуються алгоритми побудови графів видимості та будується мережа слів. В результаті проведення досліджень встановлено, що такі слова, як “uranium”, “nuclear”, “waste”, “Jupiter”, “Mercury”, “Moon”, “Earth”, “comet”, “space” та інші є ключовими для предметної області “Космічний простір”. Також у роботі порівнюються результати застосування алгоритму побудови графів видимості з алгоритмом побудови компактифікованого графу горизонтальної видимості. Досліджуючи предметну область “Комп’ютерна графіка” встановлено, що у випадку застосування алгоритму побудови компактифікованого графу горизонтальної видимості такі ключові слова, як “design”, “graphic”, “graphics”, “display”, “tiff” мають більше зв’язків у мережі, ніж у випадку застосування алгоритму побудови графів видимості. В якості допоміжних інструментів для дослідження використовуються пакет візуалізації та моделювання графів Gephi та власний набір спеціально розроблених модулів на Python. Запропонований метод може бути використаний для візуалізації певної предметної області, а також в системах інформаційної підтримки автоматизації процесів прийняття рішень, даючи змогу виявити найбільш важливі компоненти предметної області. Також результати роботи можуть бути використані під час створення персональних пошукових інтерфейсів користувачів інформаційно-пошукових систем, що, в свою чергу, дозволить спростити процес пошуку необхідної інформації.

Ключові слова: масив документів, предметна область, часовий ряд, мережа слів, статистична вага слова, граф видимості, компактифікований граф горизонтальної видимості.

Вступ. Розвиток інформаційних ресурсів в мережі Інтернет спричинив ряд специфічних проблем, пов’язаних, в першу чергу, зі стрімким збільшенням обсягів даних у веб-просторі, зокрема, і непотрібних, шумових. При цьому виявилось, що багато задач, що виникають під час роботи з мережевим інформаційним простором [1], мають багато чого спільного з математичними науками. В останні роки все більшу популярність отримала область дискретної математики, що має назву теорія складних мереж [2], яка вивчає характеристики мереж, враховуючи не тільки їх топологію, але й статистичні явища, розподіл вагових значень окремих вузлів та ребер, ефекти протікання і провідності в таких мережах струму, рідини, інформації і т. д. Цей факт відкриває широкі можливості для застосування потужного математичного апарату [1], [3]. Враховуючи проблеми розмірності та динаміки інформаційних ресурсів в глобальних мережах, для дослідження інформаційних потоків

застосовується знання з області дискретної математики (теорії графів, мереж), розпізнавання образів (класифікація, кластерний аналіз), лінгвістики, цифрової обробки сигналів, вейвлет і фрактального аналізу. Оскільки в інформаційних сховищах, розподілених у мережі, накопичуються терабайти текстових даних, то для забезпечення пошуку розміщеної в мережі інформації виникає потреба у розробленні нових підходів та методів дослідження цих даних. При цьому, безумовно, повинні враховуватись переваги та недоліки вже існуючих моделей та алгоритмів інформаційного пошуку. Більшість пошукових машин використовують класичні моделі пошуку [3]. В межах цих моделей документи розглядаються як набори ключових слів, що зустрічаються у цих документах. Незважаючи на ряд недоліків, досить популярними в застосуванні є класичні теоретико-множинні моделі. Причиною цього є їхня простота. Ймовірнісні моделі пропонують найбільш природний спосіб формально описати проблему інформаційного пошуку, але їхня популярність відносно невелика. Найбільш популярними є алгебраїчні моделі, оскільки їхня практична ефективність зазвичай виявляється вищою. Тож нові моделі інформаційного пошуку найчастіше є гібридними і мають властивості моделей різних класів.

Сучасний розвиток технологій дозволяє у деяких випадках знаходити необхідну інформацію в мережах. Але досі залишаються невирішеними проблеми подальшого аналітичного оброблення 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 завдання семантичної обробки надвеликих динамічних текстових масивів інформації.

Аналіз останніх досліджень і публікацій. Тематика цього дослідження широко зустрічається в роботах зарубіжних та вітчизняних науковців. Так, наприклад, у роботах [4], [5] акцентується увага на розробці нових методів та алгоритмів, призначених для обробки надвеликих текстових масивів. У роботах [6], [7] розглядається лінгвістичне оброблення природномовних текстів як одна з центральних проблем інтелектуалізації інформаційних технологій. Зокрема, у роботах [8] - [11] запропоновано алгоритм побудови графів видимості – VG (Visibility Graphs). Також метод створення мереж слів з використанням графів видимості представлений у роботах [12] - [16].

Метою даної роботи є розроблення методу створення мереж слів на основі текстів, використовуючи алгоритми побудови графів видимості.

Виклад основного матеріалу дослідження. У даній роботі для масиву текстових документів будується мережа зв'язків між термінами та поняттями, що входять до даних документів. Створення мереж слів, вузлами яких є елементи тексту, дає змогу виявити структурні елементи без яких цей текст втрачає свою зв'язність. При цьому актуальною є задача визначення того, які із ключових структурних елементів виявляться також інформаційно-важливими, тобто такими, що визначають структуру тексту.

Існує декілька підходів до створення мереж із текстів, так званих мереж слів (Language Network), і безліч способів інтерпретації вузлів та зв'язків, що призводить до різних видів представлення таких мереж. Вузли можуть бути з'єднані між собою, якщо відповідні їм слова знаходяться поруч у тексті [17], [18], належать одному реченню або абзацу [19], поєднані синтаксично [20], [21] або семантично [22], [23].

Для створення мережі слів в даній роботі використовується алгоритм побудови графів видимості [8]. Цей алгоритм ставить у відповідність часовому ряду граф, сформований з його елементів. Наприклад, для часового ряду {0.125, 0.063, 0.042, 0.104, 00.125, 0.063, 0.042, 0.104, 0.125, 0.063, 0.042, 0.104}, отриманий граф видимості представлений на рис. 1, де кожному вузлу в тому ж порядку відповідає елемент часового ряду.

Між вузлами, які відповідають елементам часового ряду, існує зв'язок, якщо вони знаходяться в "прямій видимості", тобто якщо їх можна з'єднати прямою лінією, що не перетинає ніяку іншу вертикальну лінію. Більш формально критерій видимості описується

наступним чином: два довільні значення (t_a, y_a) та (t_b, y_b) матимуть видимість, а отже, є двома зв'язаними вузлами відповідного графа, якщо значення (t_c, y_c), що знаходиться між ними задовольняє умову:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}.$$

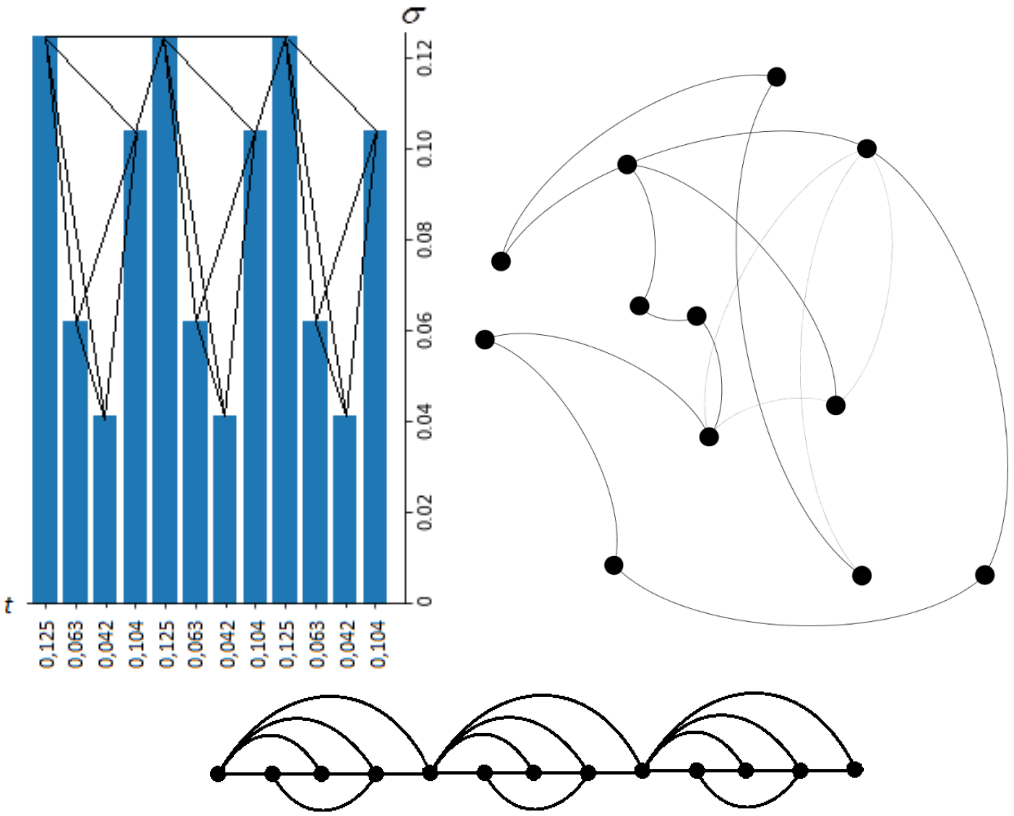

Рисунок 1 – Приклад часового ряду та відповідного графу видимості

У роботі [8] показано, що структура часового ряду зберігається у топології графу: періодичний ряд трансформується у регулярний граф, випадковий ряд – у випадковий граф, фрактальний ряд – у безмасштабний граф.

У роботах [12], [13], [14] запропоновано алгоритм побудови мереж слів – алгоритм побудови компактифікованого графу горизонтальної видимості (Compactified Horizontal Visibility Graph – CHVG). Загалом, мережа слів з використанням алгоритму горизонтальної видимості будується у три етапи. На першому етапі на горизонтальній осі відмічається ряд вузлів, кожен з яких відповідає словам у тому порядку, в якому вони з'являються в тексті, а по вертикальній осі відкладаються вагові значення – числові оцінки. На другому етапі будується граф горизонтальної видимості. Більш формально ідею побудови графу горизонтальної видимості можна представити наступним чином. Два вузли t_i і t_j, які відповідають елементам часового ряду x_i і x_j, знаходяться у горизонтальній видимості тоді й тільки тоді, коли

$$x_k < \min\{x_i, x_j\},$$

$$\text{для всіх } t_k \ (t_i < t_k < t_j).$$

Третій етап полягає в тому, що отримана на попередніх етапах мережа копмактифікується. В результаті буде отримано нову мережу слів – компактифікований граф горизонтальної видимості (CHVG) – рис. 2.

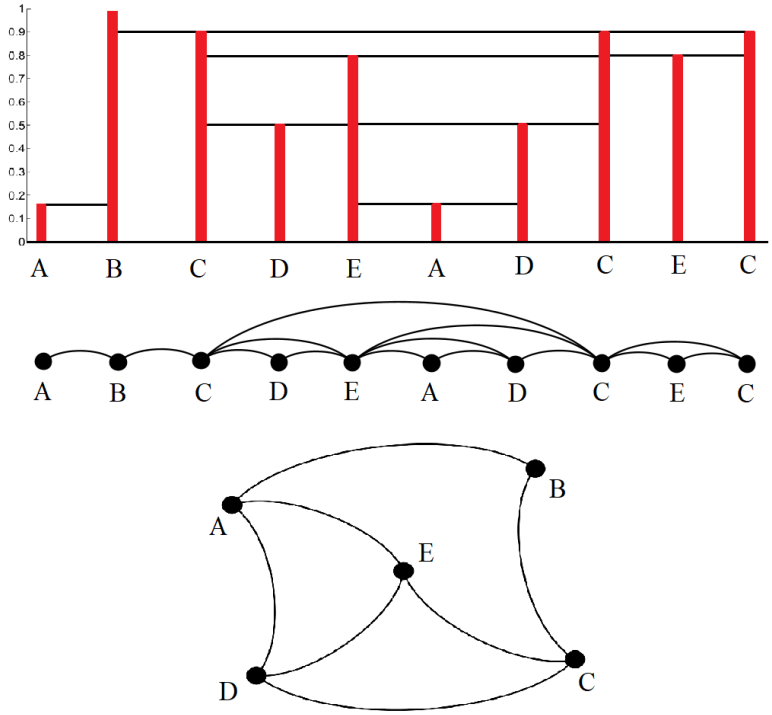

Рисунок 2 – Етапи побудови компактифікованого графу горизонтальної видимості

Таким чином, алгоритми побудови графів видимості дозволяють будувати мережеві структури на основі текстів у випадку, коли окремим словам або словосполученням поставлені у відповідність числові вагові значення. Для формування часового ряду в якості функції, яка ставить у відповідність слову число, в даній роботі використовується статистичний ваговий показник TF-IDF (з англ. Term Frequency – частота слова, Inverse Document Frequency – обернена частота документа), хоча це не єдиний можливий для вирішення завдання виділення ключових термінів підхід [13]. Цей статистичний ваговий показник використовується для оцінки важливості слів у контексті документа, що є частиною колекції документів чи корпусу [24]. Вага (значимість) слова пропорційна кількості вживань цього слова у документі і обернено пропорційна частоті вживання слова у інших документах колекції. Показник TF-IDF використовується в задачах аналізу текстів та інформаційного пошуку. Його можна застосовувати як один з критеріїв релевантності документа до пошукового запиту [25].

TF – відношення числа входжень обраного слова до кількості слів у документі. Таким чином, оцінюється важливість слова t_i в межах обраного документа. Термін введений Карен Спарк Джонс [26].

$$TF = \frac{n_i}{\sum_k n_k},$$

де n_i – число входжень слова в документ;

$\sum_k n_k$ – загальна кількість слів у документі.

IDF – інверсія частоти, з якою слово зустрічається в документах колекції. Використання IDF зменшує вагу широковживаних слів.

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|},$$

де $|D|$ – кількість документів колекції;

$\left|(d_i \supset t_i)\right|$ – кількість документів, в яких зустрічається слово t_i (коли $n_i \neq 0$).

Вибір основи логарифму у формулі не має значення, адже зміна основи призведе до зміни ваги кожного слова на постійний множник, тобто вагове співвідношення залишиться незмінним. Іншими словами, показник TF-IDF – це добуток двох множників TF та IDF

$$\text{TF-IDF} = \text{TF} \circ \text{IDF}.$$

Більшу вагу TF-IDF отримають слова з високою частотою появи в межах документа та низькою частотою вживання в інших документах колекції.

Після представлення документів колекції у вигляді числових векторів, що відображають важливість використання кожного слова з деякого набору слів (кількість слів набору визначає розмірність вектора) в кожному документі, застосовується вищеописаний алгоритм побудови графів видимості. Також, перед використанням алгоритму пропонується попередньо вилучити стоп-слова, які не мають ніякого смислового навантаження, тобто є інформаційно-неважливими.

Для проведення досліджень використано масив заздалегідь вибраних текстових документів, що тематично пов'язані з актуальною предметною областю – космічний простір. Все частіше на МКС запускають нову партію матеріалів для досліджень і наукових експериментів. Все більше приватних компаній із дослідження космосу здійснюють запуск комерційних космічних польотів, які стали уже чимось значно більшим, ніж бізнес ідеєю. Також актуальним є питання масштабного забруднення космосу. Концентрація космічного сміття на орбіті Землі досягла такої щільності, що безпечне розміщення чергового супутника на орбіті є непростим завданням.

В результаті застосування запропонованого методу створення мереж слів на основі текстів з використанням графів видимості отримано мережу (див. рис. 3) із ключових слів, які є найбільш важливими компонентами розглянутої предметної області. Для візуалізації отриманих результатів в даній роботі використовуються власний набір спеціально розроблених модулів на Python та пакет візуалізації та моделювання графів Gephi.

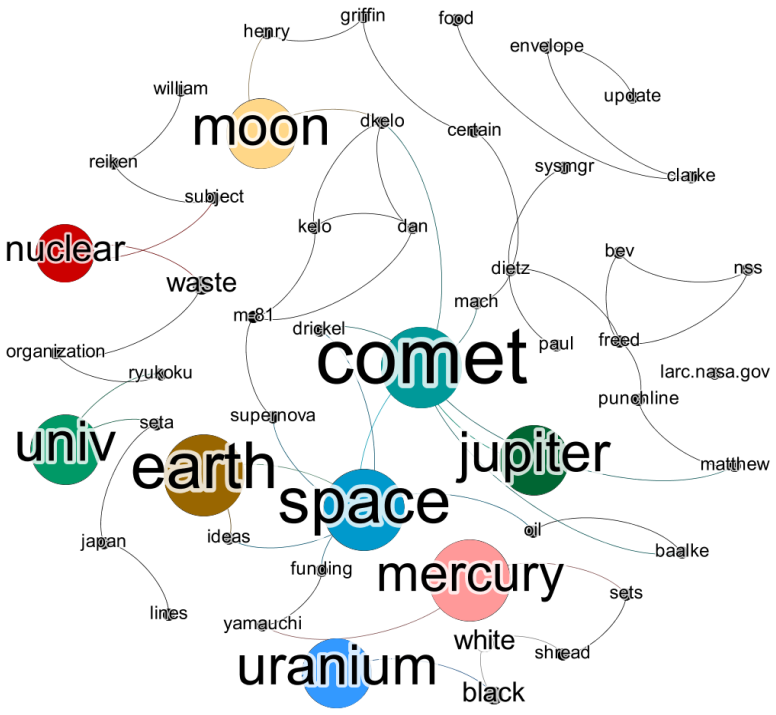

Рисунок 3 – Мережа ключових слів предметної області "Космічний простір"

Аналізуючи отримані результати (див. табл. 1) можна зробити висновок, що у випадку використання документів, що описують одну предметну область, доцільніше застосовувати

лише статистичний показник TF, а не TF-IDF. Це можна пояснити тим, що слова, які є ключовими для розглянутої предметної області й зустрічаються у більшості документів, матимуть низьке числове значення IDF (отже, і низьким буде числове значення TF-IDF), в той час, коли насправді ці слова є інформаційно-важливими, тобто такими, що визначають структуру тексту.

Таблиця 1 – ТОП-50 слів з найбільшим значенням статистичного показника TF-IDF та TF для набору документів, що тематично пов'язані з космічним простором

Weight (TF-IDF)	Word	Weight (TF)	Word
0.316867	dietz	0.135135	freed
0.298174	shread	0.114286	dietz
0.247937	szabo	0.103774	white
0.241326	ryukoku	0.093023	funding
0.241326	seta	0.084906	black
0.240817	white	0.074074	sets
0.235408	black	0.074074	shread
0.22785	sysmgr	0.071823	uranium
0.224736	funding	0.071429	matthew
0.223303	uranium	0.070175	food
0.221626	drickel	0.068493	szabo
0.217587	kelo	0.068182	oil
0.217587	dan	0.066667	japan
0.217587	dkelo	0.066667	subject
0.216796	sets	0.066667	nuclear
0.215644	c.o.egalon	0.066667	ryukoku
0.215644	punchline	0.066667	waste
0.209053	matthew	0.066667	update
0.203937	clarke	0.066667	william
0.201268	terraforming	0.066667	organization
0.201105	mach	0.066667	seta
0.19955	oil	0.066667	lines
0.19567	nss	0.066667	reiken
0.19567	supernova	0.064103	henry
0.194775	smilor	0.06383	jupiter
0.193922	larc.nasa.gov	0.063291	mercury
0.188688	lindgren	0.061224	drickel
0.187226	sean	0.057471	moon
0.187226	gallas2	0.057143	paul
0.187226	gallagher	0.056604	sysmgr
0.185197	food	0.056338	clarke
0.184839	univ	0.056338	envelope
0.184839	reiken	0.056075	baalke
0.184839	update	0.055556	mach
0.180119	bev	0.055556	comet
0.178905	roland	0.054054	nss
0.177718	yamauchi	0.054054	kelo

Продовження Таблиці 1

Weight (TF-IDF)	Word	Weight (TF)	Word
0.175723	pyron	0.054054	supernova
0.175158	envelope	0.054054	dkelo
0.170085	shell	0.054054	m-81
0.161061	william	0.054054	bev
0.158895	commerical	0.053571	punchline
0.158895	bend	0.053571	c.o.egalon
0.154707	japan	0.053333	yamauchi
0.154038	seale	0.052632	certain
0.149087	lincoln	0.051724	earth
0.149087	uaf/space	0.051724	space
0.145926	certain	0.051724	ideas
0.143763	oliveira	0.0506329	griffin

Для порівняння результатів застосування алгоритму побудови графів видимості та алгоритму побудови компактифікованого графу горизонтальної видимості використано колекцію текстових документів, що тематично пов’язані зі ще однією актуальною предметною областю – комп’ютерна графіка. Комп'ютерна графіка має багато напрямків застосування. За допомогою комп’ютерної графіки вирішують багато графічних задач. У комп'ютері синтезуються прості і складні об'єкти: поверхні, тіла, структури. Без швидкого і точного рішення графічних і геометричних задач важко уявити процес освоєння космосу, конструювання складних механізмів і машин, будову інженерних споруд, розвиток медицини і т. ін. Комп'ютерну графіку широко застосовують при рішенні актуальної проблеми підвищення продуктивності і точності інженерної роботи. Цього досягають автоматизацією розрахунково-графічних робіт, вирішуючи різноманітні задачі в області машинобудування, літакобудування і т. ін.

В результаті застосування алгоритмів видимості отримано дві різні мережі (див. рис. 4 та 5) із ключових слів, які є найбільш важливими компонентами розглянутої предметної області.

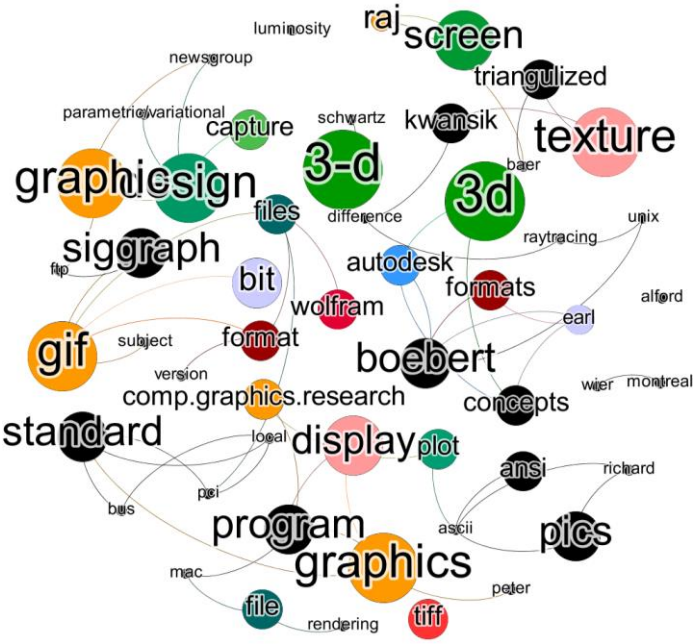

Рисунок 4 – Мережа ключових слів предметної області “Комп’ютерна графіка” отримана за алгоритмом побудови графів видимості

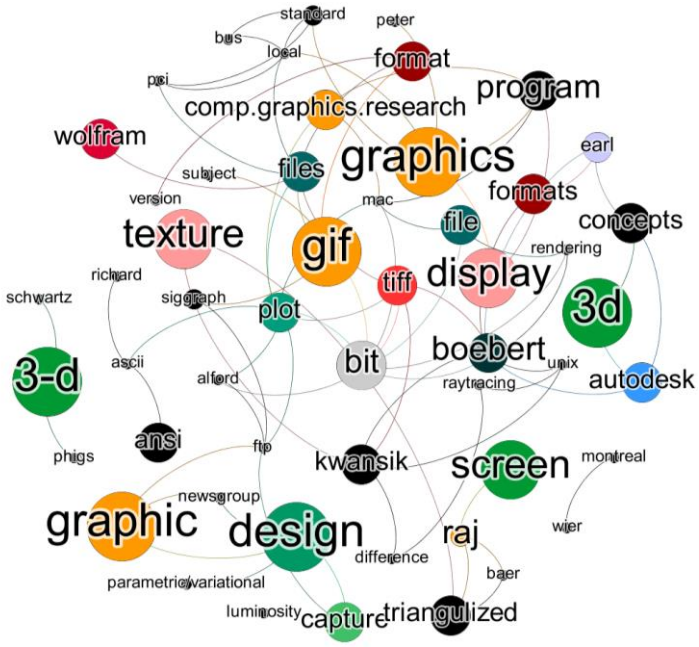

Рисунок 5 – Мережа ключових слів отримана за алгоритмом горизонтальної видимості

Таблиця 2 – ТОП-50 слів з найбільшим значенням статистичного показника TF для набору документів, що тематично пов’язані з комп’ютерною графікою

Weight (TF)	Word	Weight (TF)	Word
0.122449	files	0.076923	graphic
0.111111	kwansik	0.076923	triangulized
0.111111	plot	0.076923	newsgroup
0.107143	ascii	0.076923	display
0.107143	boebert	0.075949	baer
0.1	formats	0.075758	wier
0.095238	rendering	0.075758	3-d
0.095238	local	0.075758	screen
0.095238	alford	0.075758	ftp
0.095057	luminosity	0.074074	raytracing
0.09375	capture	0.071429	concepts
0.089286	wolfram	0.071429	autodesk
0.085714	parametric/variational	0.071429	comp.graphics.research
0.085366	bit	0.071429	gif
0.083333	version	0.071429	3d
0.078947	mac	0.071429	format
0.078652	peter	0.071429	bus
0.078431	unix	0.071429	siggraph
0.077465	file	0.071429	standard
0.076923	raj	0.071429	earl
0.076923	design	0.071429	ansi
0.076923	subject	0.071429	pci
0.076923	texture	0.071429	montreal
0.076923	program	0.071429	richard
0.076923	graphics	0.070588	tiff

Порівнюючи результати, можна помітити, що у випадку застосування алгоритму побудови компактифікованого графу горизонтальної видимості (див. рис. 5) такі ключові слова, як “design”, “graphic”, “graphics”, “display”, “tiff” мають більше зв’язків у мережі, ніж у випадку застосування алгоритму побудови графів видимості (див. рис. 4).

Висновки. У даній роботі запропонований метод створення мереж слів (Language Network). Із масиву текстових документів, тематично пов’язаних з космічним простором, виділені окремі слова та ключові поняття. Використовуючи статистичний показник TF-IDF окремим словам поставлені у відповідність числові вагові значення, і як результат, сформований часовий ряд. Використовуючи алгоритм побудови графів видимості як інструмент для аналізу часових рядів, між отриманими ключовими поняттями побудовано граф предметної області “Комп’ютерна графіка”. Аналізуючи отримані результати дослідження знайдено важливі структурні елементи, такі як “uranium”, “nuclear”, “waste”, “Jupiter”, “Mercury”, “Moon”, “Earth”, “comet”, “space”, які також виявилися інформаційно-важливими, тобто такими, що визначають структуру текстових документів пов’язаних з досліджуваною предметною областю. Показано, що у випадку використання документів, що описують одну предметну область, доцільніше застосовувати лише статистичний показник TF, а не TF-IDF.

Також порівняно результати застосування алгоритму побудови графів видимості з алгоритмом побудови компактифікованого графу горизонтальної видимості. Як результат, встановлено, що у випадку застосування алгоритму побудови компактифікованого графу горизонтальної видимості такі ключові слова, як “design”, “graphic”, “graphics”, “display”, “tiff” мають більше зв’язків у мережі, ніж у випадку застосування алгоритму побудови графів видимості.

Запропонований метод може бути використаний для візуалізації певної предметної області, а також в системах інформаційної підтримки автоматизації процесів прийняття рішень, даючи змогу виявити найбільш важливі компоненти предметної області та становити зв’язки міжними. Також результати роботи можуть використані під час створення персональних пошукових інтерфейсів користувачів інформаційно-пошукових систем, що, в свою чергу, дозволить спростити процес пошуку необхідної інформації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Д. В. Ландэ, А. А. Снарский, и И. В. Безсуднов, *Интернетика: Навигация в сложных сетях: модели и алгоритмы*. Москва, Российская Федерация: Editorial URSS, 2009.

[2] M. E. J. Newman, “The structure and function of complex networks”, *SIAM Review*, vol. 45. pp. 167-256, 2003.
doi: 10.1137/S003614450342480.

[3] Д. В. Ландэ, *Поиск знаний в Internet. Профессиональная работа*. Москва, Российская Федерация: “Вильямс”, 2005.

[4] C. C. Aggarwal, and C. X. Zhai, “Mining text data”, *Springer Science & Business Media*, pp. 77-128, 2012.
doi: 10.1007/978-1-4614-3223-4_1.

[5] G. Miner, J. Elder IV, and T. Hill, *Practical text mining and statistical analysis for non-structured text data applications*, Waltham, USA: Academic Press, 2012.
doi: 10.1016/C2010-0-66188-8.

[6] В. Ю. Тарануха, Інтелектуальна обробка текстів. Київ, Україна, 2014 [Электронный ресурс]. Доступно: www.csc.knu.ua/library/books/taranukha-40.pdf.

[7] Е. И.Большакова, О. В. Пескова, Э. С. Клышинский, А. А. Носков, Д. В. Ландэ, и Е. В. Ягунова, *Автоматическая обработка текстов на естественном языке и компьютерная лингвистика*. Москва, Российская Федерация, 2012 [Электронный ресурс]. Доступно: http://www.webground.su/data/lit/bolshakova_klyshinsky_lande_noskov_peskova_yagunova/Avtomaticheskaya_obrabotka_tekstov.pdf/.

[8] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J.C. Nuño, "From time series to complex networks: the visibility graph", *Proc. Natl. Acad. Sci. USA 105,* pp. 4972-4975, 2008. doi: 10.1073/pnas.0709247105.

[9] A. Nunez, L. Lacasa, J. Gomez, and B. Luque, "Visibility algorithms: A short review, Frontiers in Graph Theory", *InTech*, pp. 119-152, 2012. doi: 10.5772/34810.

[10] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series", *Physical Review E*, no. 80(4), pp. 1-11, 2009. doi: 10.1103/PhysRevE.80.046103.

[11] G. Gutin, T. Mansour, and S. Severini, "A characterization of horizontal visibility graphs and combinatoris on words", *Physica A*, vol. 390, iss. 12, pp 2421-2428, 2011. doi: 10.1016/j.physa.2011.02.031.

[12] D. V. Lande, and A. A. Snarskii, "Compactified HVG for the Language Network", in *Proc. of the International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science*, Chisinau, 2013, pp. 108-113.

[13] D.V. Lande, A.A. Snarskii, E.V. Yagunova, and E. Pronoza, "The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text", In: *Proceedings of the 12th Mexican International Conference on Artificial Intelligence,* 2013, pp. 209-215 doi: 10.1109/MICAI.2013.33.

[14] D. V. Lande, A. A. Snarskii, and E. V. Yagunova, "Application of the CHVG-algorithm for scientific texts', in *Proc. of the Open Semantic Technologies for Intelligent Systems (OSTIS)*, Minsk, 2014, pp. 199-204.

[15] D. V. Lande, A. A. Snarskii, and D. Yu. Manko, "The Model of Words Cumulative Influence in a Text", in *Proc. of XVIII International Conference on Data Science and Intelligent Analysis of Information*, Cham, 2018, pp. 249-256.

[16] D. Lande, A. Snarskii, E. Yagunova, E. Pronoza, and S. Volskaya, "Hierarchies of Terms on the Euromaidan Events: Networks and Respondents Perception", in *Proc. 12th International Workshop on Natural Language Processing and Cognitive Science NLPCS*, pp. 127-139, 2015.

[17] R. Ferrer-i-Cancho, and R. Solé, "The Small World of Human Language", in *Proc. of the Royal Society of London, London*, 2001, pp. 2261-2265. doi: 10.1098/rspb.2001.1800.

[18] S. N. Dorogovtsev, and J. F. Mendes, "Language as an Evolving Word Web", in *Proc. of the Royal Society of London, London*, 2001, pp. 2603-2606. doi: 10.1098/rspb.2001.1824

[19] S. Caldeira, T. Petit Lobao, R. Andrade, A. Neme, and J. Miranda, "The network of concepts in written texts", *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 49, iss. 4, pp. 523-529, 2006. doi: 10.1140/epjb/e2006-00091-3.

[20] R. Ferrer-i-Cancho, R. Solé, and R. Kohler, "Patterns in syntactic dependency networks", *Physical Review E*, vol. 69, iss. 5, pp. 051915, 2004. doi: 10.1103/PhysRevE.69.051915.

[21] R. Ferrer-i-Cancho, "The variation of Zipf's law in human language", *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 44, iss. 2, pp. 249-257, 2005. doi: 10.1140/epjb/e2005-00121-8.

[22] A. Motter, A. De Moura, Y. Lai, and P. Dasgupta, "Topology of the conceptual network of language", *Physical Review E*, vol. 65, iss. 6, pp. 1-4, 2002. doi:10.1103/PhysRevE.65.065102.

[23] M. Sigman, and G. Cecchi, "Global Organization of the Wordnet Lexicon", in *Proc. of the National Academy of Sciences*, Washington, 2002, pp.1742-1747. doi: 10.1073/pnas.022341799.

[24] J. D. Ullman, “Data Mining, Mining of massive datasets”, *Cambridge University Press*, pp. 1-17, 2011.
doi:10.1017/CBO9781139058452.002.
[25] J. Beel, B. GIPP, S. Langer, and C. Breitinger, “Research-paper recommender systems: a literature survey”, *International Journal on Digital Libraries*, vol. 17, iss. 4, pp. 305-338, 2016.
doi: 10.1007/s00799-015-0156-0.
[26] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval”, *Journal of Documentation*, vol. 28, iss.11, pp. 11-21, 2004.
doi: 10.1108/eb026526.
[27] J. M. Kleinberg, “Authoritative sources in a hyperlink environment”, *Journal of the ACM JACM*, vol. 46, iss. 5, pp. 604-632, 1999.
doi: 10.1145/324133.324140.

Стаття надійшла до редакції 29 вересня 2018 року.

REFERENCE

[1] D. V. Lande, A. A. Snarskii, and I. V. Bezsudnov, *Internetika: Navigation in complex networks: models and algorithms*. Moscow, Russia: Editorial URSS, 2009.
[2] M. E. J. Newman, “The structure and function of complex networks”, *SIAM Review*, vol. 45. pp. 167-256, 2003.
doi: 10.1137/S003614450342480.
[3] D. V. Lande, *Knowledge Search in Internet. Professional work*. Moscow, Russia: “Viliams”, 2005.
[4] C. C. Aggarwal, and C. X. Zhai, “Mining text data”, *Springer Science & Business Media*, pp. 77-128, 2012.
doi: 10.1007/978-1-4614-3223-4_1.
[5] G. Miner, J. Elder IV, and T. Hill, *Practical text mining and statistical analysis for non-structured text data applications*, Waltham, USA: Academic Press, 2012.
doi: 10.1016/C2010-0-66188-8
[6] V. Yu. Taranukha, *Intelligent processing of texts*. Kiev, Ukraine, 2014 [Online]. Available: www.csc.knu.ua/library/books/taranukha-40.pdf.
[7] E. I. Bolshakova, E. S. Klyshinsky, D. V. Lande, A. A. Noskov, O. V. Peskova, and E. V. Yagunova, *Automatic processing of texts in a natural language and computational linguistics*. Moscow, Russia, 2011 [Online]. Available: http://www.webground.su/data/lit/bolshakova_klyshinsky_lande_noskov_peskova_yagunova/Avtomaticheskaya_obrabotka_tekstov.pdf/.
[8] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J.C. Nuño, “From time series to complex networks: the visibility graph”, *Proc. Natl. Acad. Sci. USA 105*, pp. 4972-4975, 2008.
doi: 10.1073/pnas.0709247105.
[9] A. Nunez, L. Lacasa, J. Gomez, and B. Luque, “Visibility algorithms: A short review, Frontiers in Graph Theory”, *InTech*, pp. 119-152, 2012.
doi: 10.5772/34810.
[10] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, “Horizontal visibility graphs: Exact results for random time series”, *Physical Review E*, no. 80(4), pp. 1-11, 2009.
doi: 10.1103/PhysRevE.80.046103.
[11] G. Gutin, T. Mansour, and S. Severini, “A characterization of horizontal visibility graphs and combinatoris on words”, *Physica A*, vol. 390, iss. 12, pp 2421-2428, 2011.
doi: 10.1016/j.physa.2011.02.031.
[12] D. V. Lande, and A. A. Snarskii, “Compactified HVG for the Language Network”, in *Proc. of the International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science*, Chisinau, 2013, pp. 108-113.

[13] D.V. Lande, A.A. Snarskii, E.V. Yagunova, and E. Pronoza, "The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text", In: *Proceedings of the 12th Mexican International Conference on Artificial Intelligence*, 2013, pp. 209-215.
doi: 10.1109/MICAI.2013.33.

[14] D. V. Lande, A. A. Snarskii, and E. V. Yagunova, "Application of the CHVG-algorithm for scientific texts', in *Proc. of the Open Semantic Technologies for Intelligent Systems (OSTIS)*, Minsk, 2014, pp. 199-204.

[15] D. V. Lande, A. A. Snarskii, and D. Yu. Manko, "The Model of Words Cumulative Influence in a Text", in *Proc. of XVIII International Conference on Data Science and Intelligent Analysis of Information*, Cham, 2018, pp. 249-256.

[16] D. Lande, A. Snarskii, E. Yagunova, E. Pronoza, and S. Volskaya, "Hierarchies of Terms on the Euromaidan Events: Networks and Respondents Perception", in *Proc. 12th International Workshop on Natural Language Processing and Cognitive Science NLPCS*, pp. 127-139, 2015.

[17] R. Ferrer-i-Cancho, and R. Solé, "The Small World of Human Language", in *Proc. of the Royal Society of London, London*, 2001, pp. 2261-2265.
doi: 10.1098/rspb.2001.1800.

[18] S. N. Dorogovtsev, and J. F. Mendes, "Language as an Evolving Word Web", in *Proc. of the Royal Society of London, London*, 2001, pp. 2603-2606.
doi: 10.1098/rspb.2001.1824

[19] S. Caldeira, T. Petit Lobao, R. Andrade, A. Neme, and J. Miranda, "The network of concepts in written texts", *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 49, iss. 4, pp. 523-529, 2006.
doi: 10.1140/epjb/e2006-00091-3.

[20] R. Ferrer-i-Cancho, R. Solé, and R. Kohler, "Patterns in syntactic dependency networks", *Physical Review E*, vol. 69, iss. 5, pp. 051915, 2004.
doi: 10.1103/PhysRevE.69.051915.

[21] R. Ferrer-i-Cancho, "The variation of Zipf's law in human language", *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 44, iss. 2, pp. 249-257, 2005.
doi: 10.1140/epjb/e2005-00121-8.

[22] A. Motter, A. De Moura, Y. Lai, and P. Dasgupta, "Topology of the conceptual network of language", *Physical Review E*, vol. 65, iss. 6, pp. 1-4, 2002.
doi:10.1103/PhysRevE.65.065102

[23] M. Sigman, and G. Cecchi, "Global Organization of the Wordnet Lexicon", in *Proc. of the National Academy of Sciences*, Washington, 2002, pp.1742-1747.
doi: 10.1073/pnas.022341799.

[24] J. D. Ullman, "Data Mining, Mining of massive datasets", *Cambridge University Press*, pp. 1-17, 2011.
doi:10.1017/CBO9781139058452.002.

[25] J. Beel, B. GIPP, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey", *International Journal on Digital Libraries*, vol. 17, iss. 4, pp. 305-338, 2016.
doi: 10.1007/s00799-015-0156-0.

[26] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 28, iss.11, pp. 11-21, 2004.
doi: 10.1108/eb026526.

[27] J. M. Kleinberg, "Authoritative sources in a hyperlink environment", *Journal of the ACM JACM*, vol. 46, iss. 5, pp. 604-632, 1999.
doi: 10.1145/324133.324140.

ДМИТРИЙ ЛАНДЭ,
ОЛЕГ ДМИТРЕНКО

СОЗДАНИЕ СЕТЕЙ СЛОВ НА ОСНОВЕ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ГРАФОВ ВИДИМОСТИ

Предлагается метод создания сетей из текстов, так называемых сетей слов (Language Network). Из массива заранее выбранных текстовых документов, описывающих определенную предметную область, выделяются отдельные слова и ключевые понятия. Используя статистический показатель TF-IDF отдельным словам ставятся в соответствие числовые весовые значения, и как результат, формируется временной ряд. Используя алгоритмы построения графов видимости как инструмент для анализа временных рядов, между полученными ключевыми понятиями строится граф предметной области. Например, в работе рассматриваются актуальные предметные области: “Космическое пространство” и “Компьютерная графика”. Для массива заранее выбранных текстовых документов, тематически связанных с понятием космического пространства и компьютерной графики, применяются алгоритмы построения графов видимости та строится сеть слов. В результате проведения исследований установлено, что такие слова, как “uranium”, “nuclear”, “waste”, “Jupiter”, “Mercury”, “Moon”, “Earth”, “comet”, “waste”, “space” и другие являются ключевыми для предметной области "Космическое пространство". Также в работе сравниваются результаты применения алгоритма построения графов видимости с алгоритмом построения компактифицированного графу горизонтальной видимости. Исследуя предметную область “Компьютерная графика” установлено, что в случае применения алгоритма построения компактифицированного графу горизонтальной видимости такие ключевые слова, как “design”, “graphic”, “graphics”, “display”, “tiff” имеют больше связей в сети, чем в случае применения алгоритма построения графов видимости. В качестве вспомогательных инструментов для исследования используются пакет визуализации и моделирования графов Gephi и собственный набор специально разработанных модулей на Python. Предложенный метод может быть использован для визуализации определенной предметной области, а также в системах информационной поддержки автоматизации процессов принятия решений, позволяя выявить наиболее важные компоненты предметной области. Также результаты работы могут быть использованы при построении персональных поисковых интерфейсов информационно-поисковых систем, что, в свою очередь, позволит упростить процесс поиска необходимой информации.

Ключевые слова: массив документов, предметная область, временной ряд, сеть слов, статистический вес слова, граф видимости, компактифицированный граф горизонтальной видимости.

DMYTRO LANDE,
OLEH DMYTRENKO

CREATION OF LANGUAGE NETWORKS BASED ON TEXTS WITH USING VISIBILITY GRAPHS ALGORITHMS

A method to constructing language networks is proposed. Key words and concepts from the set of documents which describe some subject domain are retrieved. Numeric values are assigned to each word using a TF-IDF metric, that is intended to reflect how important a word is to a document in a collection or corpus. As the result a time series are constructed. A tool in time series analysis – the visibility graph algorithm is used for constructing the graph of subject domain. In this article two actual subject domains (“Space” and “Computer graphic”) are considered for example. The proposed method is used for the set of documents, which are related with “Space” and “Computer graphic”. A network of connections between terms and concepts, which go into textual documents is builded. Building networks of words, the nodes of which are elements of the text, enables to reveal key components of the text. At the same time, the task of determining the important

structural elements of the text which are also informationally important, is actual. As a result of the research, it was found that such words as “uranium”, “nuclear”, “waste”, “Jupiter”, “Mercury”, “Moon”, “Earth”, “comet”, “space” and others are key for the subject area “Space”. This article shows that applying only a TF metric is more expedient compared with the TF-IDF metric in case when the set of documents describe one subject domain. Also the results of applying the visibility graphs algorithm and the compactified horizontal visibility graph algorithm are compared. It was found that in some case using the compactified horizontal visibility graph algorithm gives a network of words with more quantity of connections between concepts compared with using the visibility graphs algorithm. An open-source visualization and exploration software for all kinds of graphs and networks Gephi and an original package of specially developed Python modules are used for simulation and visualization as an additional tool. The proposed method can be used for visualization some subject domain, and also for information decision support systems, enabling to reveal key components of a subject domain. Also the results of this article can be used for building UI of information retrieval systems, enabling to make a process of search a relevant information easier.

Keywords: set of documents, domain, time series, network of words, statistical weight of word, visibility graph, compactified horizontal visibility graph.

Дмитро Володимирович Ланде, доктор технічних наук, старший науковий співробітник, завідувач відділом спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

ORCID: 0000-0003-3945-1178.

E-mail: dwlande@gmail.com.

Олег Олександрович Дмитренко, аспірант, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

ORCID: 0000-0001-8501-5313.

E-mail: dmytrenko.o@gmail.com.

Дмитрий Владимирович Ландэ, доктор технических наук, старший научный сотрудник, заведующий отделом специализированных средств моделирования, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина.

Олег Александрович Дмитренко, аспирант, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина.

Dmytro Lande, doctor of technical science, senior researcher, head of the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.

Oleh Dmytrenko, postgraduate student, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.