ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ
Національна академія наук України

ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ
Національна академія наук України

Кваліфікаційна наукова
праця на правах рукопису

ДМИТРЕНКО ОЛЕГ ОЛЕКСАНДРОВИЧ

УДК 004.912

ДИСЕРТАЦІЯ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ФОРМУВАННЯ ТА АНАЛІЗУ МЕРЕЖЕВИХ МОДЕЛЕЙ ПРЕДМЕТНИХ ГАЛУЗЕЙ НА ОСНОВІ ЛІНГВОСТАТИСТИЧНОГО ПІДХОДУ

122 «Комп’ютерні науки»
Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ О. О. Дмитренко

Наукові керівники:
ЛАНДЕ Дмитро Володимирович,
доктор технічних наук,
професор
ЦИГАНОК Віталій Володимирович,
доктор технічних наук,
с.н.с.

Київ – 2024

АНОТАЦІЯ

Дмитренко О.О. Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки». – Інститут проблем реєстрації інформації, НАН України, Київ, 2024.

У дисертаційній роботі представлені результати проведених здобувачем досліджень, які виконують актуальне наукове завдання формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу обробки тематичних текстових даних та інформаційних потоків.

Актуальність цього дослідження пов'язана з тим, що з початком стрімкого розвитку інформаційно-комунікаційних технологій та глобалізацією інформаційного простору розпочалося стрімке збільшення інформаційних ресурсів, що розподілені у вебмережі. Їх розвиток відбувається куди швидше, ніж коли-небуть раніше. І як наслідок, це призвело до збільшення динамічних інформаційних потоків і, відповідно, стрімкого збільшенням об'ємів даних, представлених у електронному вигляді. Важливо зазначити й той факт, що обсяг вищезгаданих даних подвоюється приблизно кожні 18 місяців. І наразі у глобальній мережі Інтернет налічується більше сотні трильйонів документів, і частина з них – це величезні масиви текстових даних, аналіз яких може дати критично важливу інформацію.

Та з експоненційним збільшенням інформаційних потоків зростає і частка неструктурованих або слабоструктурованих даних, що, безперечно, ускладнює пошук необхідної та релевантної інформації. Наприклад, основна частина таких даних (близько 95%) є неструктурованими, і лише зовсім мала (близько 5%) – це різні бази даних, де зберігається структурована інформація, яка може бути використана під час прийняття рішень. Тож перед інформаційним суспільством постає також і ряд специфічних проблем, пов'язаних зокрема з критичною

невідповідністю між розвитком сучасних інформаційних систем і збільшенням динамічних інформаційних потоків у глобальних комп’ютерних мережах. А тому питання подальшої комп’ютеризованої обробки текстових даних з метою екстрагування знань та подальшої їх структуризації у вигляді певної онтології є важливим та актуальним у сучасному інформаційному середовищі.

Метою дисертаційної роботи є розробити нові методи побудови мережевих моделей предметних галузей на основі текстових корпусів і лінгвостатистичного аналізу текстів та розробити нові методи аналізу сформованих мереж для того, щоб приймати ефективні рішення у відповідних предметних галузях, з якими змістовно пов’язані тексти. Дисертаційна робота спрямована на вдосконалення і розширення існуючих підходів до моделювання мережевої структури предметних галузей на основі лінгвістичних даних. Основні завдання включають розробку алгоритмів побудови мереж, які враховують специфіку текстових даних, а також розробку методів аналізу отриманих мереж з метою виявлення ключових зв'язків та характеристик, які допоможуть у прийнятті обґрунтованих рішень у відповідних галузях заснованих на досліджуваних текстових даних. Об’єктом дослідження є процес структуризації у вигляді мережевих моделей текстових інформаційних потоків, розподілених у вебмережі. Предметом дослідження є лінгвостатистичні методи побудови та аналізу мережевих моделей предметних галузей на основі текстових корпусів.

Для вирішення проблеми та поставлених задач для досягнення мети були використані наступні наукові методи: методи автоматичної обробки та аналізу природної мови та методи комп’ютерної лінгвістики, завдяки яким проводилась попередня комп'ютеризована обробка природномовних текстів, лексичний аналіз та виявлення семантичних зв'язків; методи статистичного аналізу, що застосовувались для виокремлення ключових термінів (слів та словосполучень) із текстових даних; та методи дискретної математики, зокрема, методи теорії графів та складних мереж, завдяки яким здійснювалась побудова мережевих моделей предметних галузей та подальше дослідження й аналіз отриманих моделей.

У дисертаційній роботі проведено огляд та аналіз сучасних лінгвостатистичних методів, що застосовуються для структуризації текстових даних шляхом побудови мережевих моделей предметних галузей. Також описано методи та алгоритми аналізу мережевих структур і підходи до комп'ютеризованої обробки та аналізу текстових документів. Крім цього було акцентовано увагу й на проблемах, які можуть виникати під час використання методів статистичного зважування. Детально розглянуто основні рівні лінгвістичної обробки тестових даних. Розглянуто основні ідеї семантичного пошуку, як одного із найперспективніших видів автоматизованого повнотекстового інформаційного пошуку.

В результаті дисертаційного дослідження вирішено актуальне науково-практичне завдання, що стосується концептуалізації та подальшої формалізації у вигляді мережі термінів неструктурованих текстових даних, що містяться у тематичних інформаційних потоках. Загалом, мережа термінів – це семантична модель представлення текстових даних, де вузли відповідають окремим ключовим термінам тексту (словам та словосполученням), які використовуються як назви концептів певної предметної галузі, з якою змістовно пов’язані тексти, а ребра – семантико-семантичним зв’язкам між ключовими термінами.

Вперше запропоновано метод виокремлення ключових термінів із текстового корпусу зі застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Запропоновано та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency, або українською – глобальна частота терміна), який на відміну від звичайного статистичного показника TF-IDF дозволяє більш ефективно знаходити ключові та інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми. Також пропонується лінгвостатистичний метод автоматичного екстрагування, дослідження динаміки та виявлення взаємозв’язків фразеологізмів в інформаційних потоках для подальшого виявлення наративів. Запропоновано форму візуального відображення інформаційного потоку в розрізі фразеологізмів – Phrasem-Diagram (Ph-Di) діаграму.

Для побудови ненаправленої мережі із ключових термінів застосовується алгоритм побудови графа горизонтальної видимості для часових рядів (Horizontal Visibility Graph algorithm – HVG). Цей підхід також дозволяє будувати мережеві структури на основі текстів, в яких окремим словам або словосполученням деяким спеціальним чином поставлені у відповідність числові вагові значення.

У цій роботі були розроблені нові правила та методи визначення напрямків зв'язків у мережі термінів, що базуються на обробці природної мови. Також розроблено підхід до визначення вагових значень зв'язків у мережі термінів та методику побудови направлених зважених мереж термінів, як семантичних моделей предметних галузей.

Також у дисертаційній роботі представлена цілісна технологічна схема формування мережевих моделей предметних галузей на основі текстових корпусів та запропоновано методологію використання мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій.

Додатково, у цій дисертаційній роботі описано методику порівняння текстових документів та визначення ступеня подібності (та розбіжності) між ними, що базується на побудові та порівнянні відповідних семантичних мереж. Під час порівняння семантичних мереж, що відповідають текстовим документам, застосовується загальноприйнятий підхід, який полягає у наступному. Розглядається різниця матриць, що відповідають цим семантичним мережам і оцінюється її норма, як міра розбіжності. У цій дисертаційній роботі використовується норма Фробеніуса, що дорівнює кореню квадратному із суми квадратів всіх елементів відповідної різниці матриць.

Також були розроблені та представлені модель середовища семантичного інформаційного пошуку, що базується на використанні методики побудови направленої зваженої мережі термінів. Також запропоновано та представлено модель ранжування як окремих документів, так і джерел інформації, що стосуються визначеної у інформаційному запиті проблемної галузі. Також наведений приклад

формування рейтингу джерел, щодо відповідності тематики інформаційного запиту на основі порівняння семантичних мереж.

Результати дослідження можуть бути використані у різних галузях, оскільки полегшують роботу з текстовою інформацією, допомагаючи у кращому розумінні та аналізі великих обсягів даних. Мережі термінів мають потенціал для розвитку й розробки ефективних систем аналізу інформації, покращення систем інформаційного пошуку, рекомендаційних систем та систем підтримки прийняття рішень, підвищуючи точність результатів пошуку та рекомендацій через краще розуміння семантики користувацьких запитів.

Представлена у цій дисертаційній роботі методика порівняння текстових документів та визначення ступеня подібності (або розбіжності) між ними, що базується на порівнянні відповідних їм семантичних мереж може бути використана для розвитку нових алгоритмів семантичного аналізу текстів і покращення процесів автоматичної обробки та порівняння текстових даних. Зокрема, представлена мережева методика порівняння текстових документів може бути використана для виявлення структурних і термінологічних розбіжностей у правовій сфері, що сприятиме парламентському контролю та гармонізації міжнарожного права.

Мережеві моделі текстів можуть бути важливим інструментом для створення, підтримки та управління базами знань у різних організаціях, допомагаючи відфільтрувати та структурувати великі обсяги інформації. Здобуті результати досліджень можуть допомогти у подальшому розвитку технологій обробки природної мови та систем автоматичної обробки текстів, що сприятиме автоматизації обробки, структуризації й аналізу текстової інформації.

Ключові слова: інформаційні системи, великі дані, база знань, обробка природної мови, комп’ютерно-лінгвістичні методи аналізу, онтологія, мережева семантична модель ключових термінів, мережевий аналіз, пошук семантичної подібності, виявлення наративів, прийняття рішень.

ABSTRACT

Dmytrenko O.O. Information technologies for formation and analysis of network models of subject domains based on linguostatistical approach. – Qualifying scientific work on the manuscript rights.

Doctor of Philosophy dissertation under 122 «Computer Science» specialty. – Institute for Information Recording of the National Academy of Sciences of Ukraine, Kyiv, Ukraine, 2024.

In the dissertation, the author presents the results of conducted research aimed at addressing the current scientific task of forming and analyzing network models of subject domains based on linguostatistical methods for processing thematic textual data and information flows.

The relevance of this research is associated with the rapid development of information and communication technologies and the globalization of the information space, which has led to a significant increase in informational resources distributed across the web. Their development is occurring much faster than ever before. As a result, this has led to an increase in dynamic information flows and, consequently, a rapid increase in the volumes of data presented in electronic form. It is important to note the fact that the volume of the aforementioned data roughly doubles approximately every 18 months. Currently, the global Internet comprises over a hundred trillion documents, and a portion of these are massive collections of text data, the analysis of which can provide critically important information.

But with the exponential increase in information flows, the proportion of unstructured or weakly structured data is growing, undoubtedly complicating the search for necessary and relevant information. For instance, the major part of such data (approximately 95%) is unstructured, with only a very small percentage (about 5%) being various databases containing structured information that can be used in decision-making processes. Therefore, the information society also faces a series of specific problems, particularly related to the critical discrepancy between the development of modern information systems and the increase in dynamic information flows within global

computer networks. Therefore, the issue of further computerized processing of text data for knowledge extraction and subsequent structuring in the form of a certain ontology remains important and relevant in the modern information environment.

The purpose of the dissertation work is to develop new methods for the construction network models of subject domains based on text corpora and linguostatistical analysis of texts and to devise new methods for analyzing the formed networks to make effective decisions within the respective subject domains that are semantically related to the texts. The dissertation aims to improve and expand existing approaches to modelling the network structure of subject domains based on linguistic data. The main tasks include developing algorithms for the construction networks that consider the specifics of text data, as well as developing methods to analyze the resulting networks to identify key connections and characteristics, aiding in making informed decisions within respective domains based on the researched text data. The object of the research is the process of structuring in the form of network models of text information flows distributed across the web. The subject of the research is linguostatistical methods for the construction and analyzing network models of subject domains based on the text corpora.

To solve the problem and achieve the set goals, the following scientific methods were used: natural language processing and computational linguistics methods were employed for preliminary computerized processing of natural language texts, lexical analysis, and identification of semantic relationships; statistical analysis methods were applied to extract key terms (words and phrases) from text data; and discrete mathematics methods, particularly graph theory and complex networks, were used for the construction network models of subject domains and further research and analysis of the obtained models.

The dissertation provides a review and analysis of modern linguostatistical methods used for structuring text data through the formation of network models of subject domains. The methods and algorithms for analyzing network structures and approaches to computerized processing and analysis of text documents are also described. Additionally, attention was focused on the problems that may arise when using statistical weighting methods. The main levels of linguistic processing of text data are considered

in detail. The main concepts of semantic search, as one of the most promising types of automated full-text information search, are considered.

As a result of the dissertation research, an actual scientific-practical task related to the conceptualization and further formalization in the form of a network of terms from unstructured text data contained in thematic information flows was solved. In general, a network of terms is a semantic model representing text data, where nodes correspond to individual key terms (words and phrases) in the text used as names of concepts within a specific subject domain related to the content, and edges represent semantic-semantic relationships between these key terms.

For the first time, a method of extracting key terms from a text corpus using a broader natural language processing based on Part-of-Speech tagging is proposed. A new statistical indicator of term importance in text, called GTF (Global Term Frequency), has been proposed and researched. Unlike the commonly used statistical indicator TF-IDF, GTF enables more effective identification of key and informationally significant elements within a pre-defined text corpus on a specific topic. Additionally, a linguostatistical method for automatic extraction, analysis of dynamics, and identification of correlations among phraseological units in information flows is proposed for further narrative detection. A visual representation form of the information flow through phraseological units - the Phrasem-Diagram (Ph-Di) - is suggested.

The application of the Horizontal Visibility Graph algorithm (HVG) for time series is used to construct an undirected network from key terms. This approach also allows to construct the network structures based on texts, in which specific words or phrases are assigned numerical weight values in a particular manner.

This work introduces new rules and methods for determining connections within a network of term based on natural language processing. Additionally, an approach for determining the weighted values of connections in the network of term and a methodology for the construction directed weighted networks of terms as semantic models of subject domains were developed.

Also, the dissertation includes a complete technological scheme for the construction network models of subject domains based on text corpora, and proposes a

methodology for using network of terms to form a knowledge base for decision support systems during information operation recognition.

Additionally, this dissertation describes a method for comparing text documents and determining the degree of similarity (and dissimilarity) between them, based on the construction and comparing respective semantic networks. During the comparison of semantic networks corresponding to text documents, a common approach is applied, which involves the following: the difference between matrices corresponding to these semantic networks is considered, and its norm is evaluated as a measure of dissimilarity. In this dissertation, the Frobenius norm is used, which equals the square root of the sum of squares of all elements of the corresponding matrix difference.

In addition, a model of the semantic information retrieval environment was also developed and presented, based on the methodology for the construction a directed weighted network of terms. Furthermore, a ranking model for both individual documents and sources of information related to a specific problem domain outlined in the information query is also proposed and presented. An example of forming a ranking of sources based on assessing the relevance of the information query's thematic correspondence through semantic networks comparison is also provided.

The results of the research be applied across various fields as they facilitate working with text information, helping in better understanding and analysis of large volumes of data. Network of terms have the potential for development and the creation of effective information analysis systems, improve information retrieval systems, recommendation systems, and decision support systems, thereby improving the accuracy of search results and recommendations through a better understanding of the semantics of user queries.

The method presented in this dissertation for comparing text documents and determining the degree of similarity (or dissimilarity) between them, based on comparing their respective semantic networks, can be used to develop new algorithms for semantic text analysis and enhance processes of automatic processing and comparison of text data. In particular, the presented network method of comparing text documents can be used to

identify structural and terminological differences in the legal field, which will contribute to parliamentary control and harmonization of international law.

Network models of texts can serve as important tools for creating, supporting, and managing knowledge bases within various organizations, helping in the filtration and structuring of large volumes of information. The obtained research results may assist in furthering the development of natural language processing technologies and automatic text processing systems, contributing to the automation, structuring, and analysis of text information.

Keywords: information systems, big data, knowledge base, natural language processing, computational linguistic analysis methods, ontology, network semantic model of key terms, network analysis, semantic similarity search, narrative detection, decision-making.

Список публікацій здобувача

1. Ланде, Д. В., & Дмитренко, О. О. (2018). Створення мереж слів на основі текстів з використанням алгоритмів графів видимості. *Information Technology and Security: Ukrainian research papers collection, 2018, Vol. 6, Iss. 2 (11).* 5-18. DOI: doi.org/10.20535/2411-1031.2018.6.2.153486.
2. Lande, D. V., Dmytrenko, O. O., & Snarskii, A. A. (2018). Transformation texts into complex network with applying visability graphs algorithms. *Информационные технологии и безопасность*. *Матеріали XVIII Міжнародної науково-практичної конференції ІТБ-2018. - К.: ООО "Инжиниринг"*, 20-33.
3. Lande, D. V., Dmytrenko, O. O., & Snarskii, A. A. (2018). Transformation texts into complex network with applying visability graphs algorithms. *Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). In CEUR Workshop Proceedings (ceur-ws.org). Vol-2318,* (pp. 95-106) *urn:nbn:de:0074-2318-4*.
4. Ланде, Д. В., Дмитренко, О. О., & Радзієвська, О. Г. (2019). Побудова онтологій в галузі права за даними сервісу Google Scholar. *Інформація і право*, *1*(4), 74-85. DOI: doi.org/10.37750/2616-6798.2019.1(28).221313
5. Ланде, Д. В., & Дмитренко, О. О. (2019). Побудова мережі термів у сфері кібербезпеки за даними сервісу Google Scholar. *Матеріали XVII Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (25 - 26 квітня 2019 р., м. Київ, Україна)*, 143-145.
6. Дмитренко О.О. (2019). Створення термінологічних онтологій предметних областей на базі ресурсу Google Scholar. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» 16-17 травня 2019 року: збірник / - Київ: ІПРІ НАН України*, 108-109.
7. Ланде, Д. В., Дмитренко, О. О., & Радзієвська, О. Г. (2019). Визначення напрямків зв’язків у мережі термінів. *Інформаційні технології та безпека.*

Матеріали XIX Міжнародної науково-практичної конференції «ІТБ-2019». Київ: ТОВ «Инжиниринг, 103-112.

8. Lande, D., Dmytrenko, O., & Radziievska, O. (2019). Determining the directions of links in undirected networks of terms. *Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019). In CEUR Workshop Proceedings (ceur-ws.org). Vol-2577*, (pp. 132-145). ISSN 1613-0073.
9. Ланде, Д. В., & Дмитренко, О. О. (2019). Визначення вагових значень зв'язків у мережі термінів. *Реєстрація, зберігання і обробка даних*, *21(4)*, 40-48. DOI: doi.org/10.35681/1560-9189.2019.21.4.199357
10. Lande, D. V., Dmytrenko, O. O., & Radziievska, O. H. (2020). Subject domain models of jurisprudence according to google scholar scientometrics data. *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference. Lviv, Ukraine, April 23-24, 2020. In CEUR Workshop Proceedings (ceur-ws.org).* - Vol-2604, (pp. 32-43). ISSN 1613-0073.
11. Ланде, Д. В., & Дмитренко, О. О. (2020). Метод побудови направлених зважених мереж термінів на основі текстових корпусів. *Матеріали XVIII Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (12 - 13 травня 2020, м. Київ, Україна)/ НТУУ "КПІ",* 68-71.
12. Ланде, Д. В., & Дмитренко, О. О. (2020). Побудова направлених зважених мереж термінів. *XI Всеукраїнська науково-практична конференція «Актуальні проблеми управління інформаційною безпекою держави»: зб. тез наук. доп. наук.-практ. конф. (Київ, 15 травня 2020 р.). [Електронне видання]. – Київ : НА СБУ*, 129-130.
13. Lande D. V., Dmytrenko O. O., Andriichuk O. V., Tsyganok V. V., & Porplenko Y. V. (2020). Building of directed weighted networks of terms for decision-making support during information operations recognition. *Математичне та імітаційне моделювання систем. МОДС 2020 : тези доповідей П'ятнадцятої міжнародної*

науково-практичної конференції (29 червня . 01 липня 2020 р., м. Чернігів) / М-во освіти і науки України ; Нац. Акад. наук України ; Академія технологічних наук України ; Інженерна академія України та ін. - Чернігів : ЧНТУ, 147-148.

14. Lande D. V., Dmytrenko O. O., Andriichuk O. V., Tsyganok V. V., & Porplenko Y. V. (2020). Building of directed weighted networks of terms for decision-making support during information operations recognition, *In: Mathematical Modeling and Simulation of Systems (MODS'2020). MODS 2020. Advances in Intelligent Systems and Computing, vol 1265*, (pp. 197-208). *Springer, Cham. Pages*. DOI: doi.org/10.1007/978-3-030-58124-4_19
15. Дмитренко О.О. (2020). Побудова мереж термінів на основі тематичних інформаційних публікацій. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» 28-29 вересня 2020 року: збірник / - Київ: ІПРІ НАН України*, 107-108.
16. Lande D. V., & Dmytrenko O. O. (2020) Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora. *In 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC*) (pp. 1-4). *IEEE.* DOI: doi.org/10.1109/SAIC51296.2020.9239182
17. Ланде, Д. В., & Дмитренко, О. О. (2020). Методика виокремлення ключових слів і словосполучень та побудови направлених зважених мереж термінів із застосуванням Part-of-Speech tagging. *Інформаційні технології і безпека. Матеріали XX Міжнародної науково-практичної конференції ІТБ-2020. - Київ: Інжиніринг,* 140-144. ISBN: 978-966-2344-77-6
18. Lande, D., Andriichuk, O., Dmytrenko, O., Tsyganok, V., & Porplenko, Y. (2020). Побудова баз знань систем підтримки прийняття рішень з використанням направлених мереж термінів при дослідженні інформаційних операцій. *Information Technology and Security*, *8*(2), 153-163. DOI: doi.org/10.20535/2411-1031.2020.8.2.222597
19. Дмитренко, О. О. (2020). Побудова направлених зважених мереж термінів із застосуванням Part-of-speech tagging. *Реєстрація, зберігання і обробка даних, 22*(4), 47-55. DOI: doi.org/10.35681/1560-9189.2020.22.4.225914

20. Lande, D. V., & Dmytrenko, O. O. (2020). Methodology for Extracting of Key Words and Phrases and Building Directed Weighted Networks of Terms with Using Part-of-speech Tagging. *Selected Papers of the XX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2020) In CEUR Workshop Proceedings (ceur-ws.org). - Vol-2859* (pp. 168-177). ISSN 1613-0073
21. Ланде, Д. В., & Дмитренко, О. О. (2021). Побудова онтологічних моделей у галузі права. *Актуальні проблеми управління інформаційною безпекою держави: зб. тез наук. доп. наук.-практ. конф. (Київ, 26 березня 2021 р.). [Електронне видання]. - Київ : НА СБУ,* 62-63.
22. Ланде, Д. В., & Дмитренко, О. О. (2021). Формалізація знань та побудова термінологічних онтологій у правовій галузі. *Парламентський контроль в умовах децентралізації державної влади та цифрової трансформації в Україні: стан і проблеми: матеріали Першої всеукраїнської науково-практичної конференції, м. Київ, 30 березня 2021 р.*, 35-39.
23. Lande D.V., & Dmytrenko O.O. (2021). Using Part-of-Speech Tagging for Building Networks of Terms in Legal Sphere. *In Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference. Kharkiv, Ukraine, April 22-23, 2021. CEUR Workshop Proceedings (ceur-ws.org). - Vol-2870* (pp. 87-97). ISSN 1613-0073.
24. Ланде, Д. В., & Дмитренко, О. О. (2021). Використання направлених зважених мереж термінів для визначення ступеня подібності текстів. *Міжнародна наукова-технічна конференція "Інтелектуальні технології лінгвістичного аналізу": Тези доповідей. - К.: НАУ*, 7.
25. Zgurovsky, M. Z., Boldak, A. O., Lande, D. V., Yefremov, K. V., Pyshnograiev, I. O., Soboliev, A. M., & Dmytrenko, O. O. (2021). *Enhancing the Relevance of Information Retrieval in Internet Media and Social Networks in Scenario Planning Tasks, IEEE International Conference on System Analysis & Intelligent Computing SAIC 2021: System Analysis & Intelligent Computing. Studies in Computational Intelligence, vol 1022. Springer, Cham*, 187-199 DOI: https://doi.org/10.1007/978-3-030-94910-5_10

26. Ланде, Д. В., & Дмитренко, О. О. (2021). Побудова семантичних мереж та визначення ступеня розбіжності текстів. *Інформація і право*, *2*(41), 44-51. DOI: doi.org/10.37750/2616-6798.2022.2(41).270362
27. Dmytrenko, O. O., & Lande, D. V. (2022). Building of semantic networks to determine the degree of text similarity or difference. *Теоретичні і прикладні проблеми фізики, математики та інформатики: матеріали XX Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених (15 червня 2022 р., м. Київ, Україна). - Київ : КПІ ім. Ігоря Сікорського, Вид. "Політехніка"*, 197-202.
28. Lande, D., Soboliev, A., & Dmytrenko, O. (2022). Intelligent technologies in information retrieval systems. *Artificial intelligence*, 27(1), 260-268. DOI: https://doi.org/10.15407/jai2022.01.260
29. Lande, D. V., Dmytrenko, O. O., Shevchenko, A. I., Klymenko, M. S., & Vakulenko, M. O. (2023). Spoken language identification based on the transcript analysis. *Digital Scholarship in the Humanities*, 38(2), 586-595. DOI: https://doi.org/10.1093/llc/fqac052
30. Дмитренко О.О. (2022). Програмний модуль автоматичного екстрагування ключових термінів з інформаційних потоків. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція 27-28 вересня 2022 року: збірник / - Київ: ІПРІ НАН України*, 122-123.
31. Zgurovsky, M. Z., Lande, D. V., Yefremov, K., Dmytrenko, O. O., Boldak, A. O., & Soboliev, A. M. (2022). Extracting and Identifying Relationships of Key Phrases in Information Flows. *In 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC) 04-07 October 2022*, (pp. 1-5). ISBN:979-8-3503-9674-4. DOI: 10.1109/SAIC57818.2022.9923019
32. Dmytrenko, O. (2022). Formation Networks of Terms for Identifying Semantic Similarity or Difference Degree of Texts in Cybersecurity. *Theoretical and Applied Cybersecurity*, 4(1). DOI: doi.org/10.20535/tacs.2664-29132022.1.274118
33. Zgurovsky, M. Z., Lande, D. V., Dmytrenko, O. O., Yefremov, K., Boldak, A. O., & Soboliev, A. M. (2022). Technological Principles of Using Media Content for

Evaluating Social Opinion. *System Analysis and Artificial Intelligence. Studies in Computational Intelligence*, Springer, Cham, 1107, 379-396 DOI: https://doi.org/10.1007/978-3-031-37450-0_22

34. Дмитренко, О. О. (2023). Формування та дослідження динамічних мереж термінів. *Інформаційні технології і безпека. Матеріали XXIII Міжнародної науково-практичної конференції ІТБ-2023.* - *Київ: Інжиніринг,* 83-84. ISBN: 978-966-2344-96-7.
35. Ланде Д. В., Дмитренко О. О., & Єфремов К. В. (2022). Комп'ютерна програма автоматичної побудови мереж термінів на основі аналізу текстових потоків «TermsNet». *Свідоцтво про реєстрацію авторського права на твір № с202204275 від 19.09.2022.* Державна організація «Український національний офіс інтелектуальної власності та інновацій».

Зміст

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NLP – Natural Language Processing

TF – Term Frequency

GTF – глобальна частота терміна (Global Term Frequency)

PoS – Part-of-Speech

VG – граф видимості (Visibility Graph)

HVG – граф горизонтальної видимості (Horizontal Visibility Graph)

HITS – Hyperlink Induced Topic Search

ІО – інформаційні операції

ІПС – інформайно-пошукова система

ІПМ – інформайно-пошукова мова

ПОД – пошуковий образ документа

СППР – система підтримки прийняття рішень

ЧКВ – часткові коефіцієнти впливу

БЗ – база знань

DWNT – Directed Weighted Network of Terms

ВСТУП

Актуальність роботи. З початком стрімкого розвитку інформаційно-комунікаційних технологій та глобалізацією інформаційного простору розпочалося стрімке збільшення інформаційних ресурсів, що розподілені у вебмережі. Сучасні інформаційно-комунікаційні технології та загалом інформаційний простір розвиваються швидше, ніж коли-небудь раніше. Як наслідок, це призвело до збільшення динамічних інформаційних потоків, а отже, й даних, які ними супроводжуються. Такий процес характеризується відповідно стрімким збільшенням об’ємів, зокрема, текстових даних, які продукуються елементами інформаційного простору, зокрема, документами та найрізноманітнішими джерелами даних – файлами, незалежно від форматів їх подання, електронними листами, вебсторінками, платформами соціальних медіа та соціальних мереж та іншими джерелами. Дані створюються, реєструються, зберігаються, обробляються та відтворюються дедалі частіше у електронному вигляді. Важливо зазначити й той факт, що обсяг вищезгаданих даних подвоюється приблизно кожні 18 місяців. Унаслідок цього за п'ять попередніх років людством було продуковано даних більше, ніж за всю попередню історію. Щодня з’являються величезні масиви зокрема текстових даних, аналіз яких може дати критично важливу інформацію. І найголовніше те, що частина джерел таких інформаційних потоків є відкритими й загальнодоступними, а отже, становлять великий інтерес практично серед всіх державних та приватних структур, в яких є необхідність в аналізі даних задля можливості швидкого прийняття рішень у сфері їх діяльності, зокрема – телекомунікаційній, кібернетичній, фінансовій, торговій, військовій, політичній, дипломатичній та інших сферах.

Та розвиток мережі Інтернет і розширення інформаційного простору спричиняють також ряд специфічних проблем, пов’язаних, в першу чергу, зі стрімким збільшенням об’ємів даних, або так званим інформаційним перевантаженням. Адже такий інформаційний сплеск, або так званий інформаційний вибух, супроводжується не лише припливом нових цінних знань. Зі

збільшенням кількості текстових інформаційних потоків зростає і частка неструктурованих або слабоструктурованих даних, в тому числі й непотрібних та шумових (так званого «інформаційного сміття»), що, безперечно, ускладнює пошук необхідної та релевантної інформації. Наприклад, основна частина таких даних (близько 95%) є неструктурованими або слабоструктурованими, і лише зовсім мала (близько 5%) – це різні бази даних, таблиці, де зберігається структурована інформація, яка може бути використана під час прийняття рішень. Виникає також проблема, що пов’язана з продукуванням дублікатів текстових документів. Тож не завжди з масивних текстових інформаційних потоків та, відповідно, величезних об’ємів текстових даних, які ними супроводжуються, користувач може отримати релевантну інформацію у відповідь на свій запит.

Тому зараз перед інформаційним суспільством постає ряд проблем, з якими ніхто раніше не стикався. Основною проблемою є критична невідповідність між розвитком сучасних інформаційних систем і експоненційним збільшенням динамічних інформаційних потоків у глобальних комп’ютерних мережах. А саме, проблема полягає у відсутності підходящих технологічних рішень та у неспроможності наявних систем обробляти величезні об’єми неструктурованих даних, зокрема текстових, й виокремлювати з них знання з тією ж самою швидкістю, з якою відповідні дані продукуються й накопичуються. Тому для забезпечення ефективного пошуку розміщеної в мережі Інтернет інформації необхідна розробка нових підходів і методів дослідження та структуризації цих даних. При цьому, безумовно, повинні враховуватись переваги та недоліки вже існуючих моделей, методів та алгоритмів.

Тож величезні об’єми текстових інформаційних потоків та динамічних текстових масивів, що пов’язані з певною проблемною предметною галуззю, обумовлюють актуальність процесу концептуалізації текстових даних, що ними супроводжуються, та їх подальшої формалізації у вигляді певної онтологічної моделі, що буде зрозуміла та придатна для обробки комп’ютером та дозволяє виконувати більш ефективну обробку та аналіз. А отже, актуальною є розробка нових та удосконалення існуючих методів та технологічних рішень, які

застосовуються для вирішення цього завдання, з метою забезпечити достатньо високу швидкість обробки й аналізу неструктурованих текстових даних, та підвищити якість та точність цих процесів. Застосування покращених методів і технологій допоможе більш ефективно виявляти ключові теми, зв'язки та залежності у текстах, що полегшить процес пошуку необхідної інформації та загалом дозволить підвищити пертинентність інформаційно-пошукової системи під час інформаційного пошуку. Розробка нових методів та технологічних рішень також сприяє автоматизації процесу обробки текстових даних, що дозволяє зекономити час та зусилля. Це особливо важливо у випадку великих обсягів даних, коли ручна обробка стає неможливою або недоцільною.

Таким чином, процес перетворення накопичених на інформаційних ресурсах неструктурованих даних у знання є ключовим у формуванні глобального інформаційного простору. Це включає в себе розробку методів та технологій для аналізу, класифікації, екстракції та інтерпретації даних з метою отримання цінних знань, які, в свою чергу, можуть бути використані для розробки рекомендаційних систем, які допомагають приймати швидкі та обґрунтовані рішення в різних галузях, таких як медицина, фінанси, маркетинг, наука та багато інших.

Отже, під час комплексних досліджень певної проблемної предметної галузі, з якою тематично пов’язані потоки текстових даних, та здійснення інформаційного пошуку важливим етапом є детальне формалізоване представлення цих даних у формі знань (набору об’єктів та сутностей реального світу та зв’язків між ними), які не лише зрозумілі людині, а й, найголовніше, стають придатними для подальшої автоматизованої обробки комп’ютером. Одним із видів такої формалізації є побудова термінологічної онтології досліджуваної предметної галузі. В якості онтології зручною й ефективною моделлю представлення текстових даних може розглядатися лінгвомережева модель – модель, що представляє собою мережу із ключових термінів (слів та словосполучень), поєднаних між собою змістовними зв’язками. Це пов’язано з тим, що, як виявилось, багато задач, які виникають під час роботи з текстовими інформаційними потоками, лежать на перетині між лінгвістикою та математичними науками, зокрема, теорією складних мереж. Цей

факт відкриває широкі можливості для застосування лінгвістичної теорії та потужного математичного апарату – теорії графів. Лінгвістична теорія як розділ загального мовознавства, в свою чергу, дає змогу працювати з природномовними текстами, знаючи при цьому контекст, структуру, властивості та будову їх елементів. З іншого боку, враховуючи проблеми, що пов’язані із розмірністю та стрімкою динамікою інформаційних ресурсів в глобальних мережах, в якості потужної математичної теорії, в межах якої може вирішуватись проблема формалізації предметної галузі, розглядаються знання з області статистики, дискретної математики, зокрема теорії графів та складних мереж. Якщо говорити у термінах теорії складних мереж, то тексти визначеної тематичної спрямованості можна представити у вигляді мережі із термінів – слів та словосполучень, пов’язаних між собою формальним смисловим зв’язком. Одним із видів такої лінгвомережевої моделі може бути семантична мережа, що сформована із ключових термінів, що зустрічаються у корпусі текстів. У ній вузли відповідають окремим ключовим поняттям предметної галузі, а направлені зв’язки та відношення – семантико-семантичним зв’язкам між ними. Семантичне представлення математично можна подати у вигляді графа та відповідної йому матриці, що відображає бінарні відношення між вузлами. Аналіз таких мереж може бути основою для прийняття рішень в обраній проблемній предметній галузі, з якою змістовно пов’язані тексти. Наприклад, порівняння матриць семантичних мереж, отриманих для різних текстів, дає змогу визначити семантичну близькість (подібність або схожість) відповідних текстів, або з іншого боку – семантичну розбіжність (невідповідність або суперечливість). Порівняння матриць семантичних мереж дозволяє визначити ступінь схожості між двома текстами: чим більше спільних зв'язків та взаємодій між словами у матриці, тим більша ймовірність, що текст має схожий семантичний зміст. Порівняння матриць семантичних мереж може бути корисним у багатьох випадках. Наприклад, в аналізі текстів, пошуку схожих документів, виявленні плагіату, класифікації текстів за їх темою або семантикою. Цей підхід дає змогу об'єктивно порівняти текстові дані, виявляти шаблони у текстах та виявити схожість між ними на основі семантичних

характеристик. Побудова мережевих моделей на основі лінгвостатистичних методів допомагає виявляти ключові теми, групувати схожі поняття та встановлювати зв'язки між ними, що дозволяє отримати більш глибоке розуміння семантичного змісту та структури текстових корпусів з метою виявлення нових знань. Також важливо зазначити, що такий підхід дає змогу використовувати лінгвомережеві моделі у процесі семантичного пошуку відповідних до запиту документів, аби забезпечити більш точні та релевантні результати: знайти відповідність між запитом користувача та документами, які містять відповідну за семантичним змістом інформацію. Під час семантичного пошуку лінгвомережеві моделі використовуються для розуміння семантичного змісту запиту та документів, що знаходяться у інформаційній базі. Лінгвомережеві моделі можуть аналізувати семантичні взаємозв'язки між словами, розпізнавати синоніми, антоніми, контекстуальні значення та інші лінгвістичні особливості. Це дозволяє забезпечити більш точний та релевантний пошук, оскільки лінгвомережеві моделі можуть враховувати не тільки слова, а й їх семантичні зв'язки, контекст, асоціації та інші фактори, які впливають на семантику тексту. Це підвищує пертинентність результатів пошуку, покращує ефективність та задоволення користувача в процесі інформаційного пошуку.

Тож отримані лігвомережеві моделі можуть бути використані в багатьох галузях для різних цілей, таких як порівняння текстових даних з метою виявлення схожого семантичного змісту, класифікація текстів за їх темою або семантикою, виявлення плагіату, виділення найважливіших фраз, наративів та ідей з тексту, формування анотацій, коротких висновків і рефератів на основі текстової інформації, розробка інформаційно-пошукових систем, класифікації текстів за їх темою або семантикою, виявлення плагіату, побудова рекомендаційних систем, аналіз текстових даних та багато інших. Вони дозволяють отримати структуровану та семантично багатошарову репрезентацію текстової інформації, що полегшує подальший аналіз та використання даних в різних галузях, таких як медицина, фінанси, маркетинг, наука, телекомунікації, кібернетична, військова, політична галузі та багато інших.

Під час побудови мереж із термінів (або просто – мереж термінів), як семантичних моделей представлення текстових даних, відкритою та до кінця не вирішеною проблемою є визначення та виокремлення базових об’єктів (ключових термінів – слів та словосполучень). Також у зв'язку зі складністю природної мови, не менш складною й відкритою проблемою концептуалізації є встановлення семантико-синтаксичних зв’язків між вузлами мережі, що відповідають термінам, визначення напрямків та вагових значень цих зв’язків. Не менш важливою є автоматизація вищезгаданих процесів.

Розглянуті в цій дисертаційній роботі лінгвостатистичні методи побудови мережевих моделей предметних галузей на основі текстових корпусів відкривають широкий спектр можливостей для автоматичної обробки великих обсягів текстової інформації з метою її подальшого аналізу й отримання цінних знань із текстів, щоб приймати рішення у проблемній предметній галузі, з якою ці тексти тематично пов’язані.

Крім цього був проведений загальний огляд вже існуючих методів, що використовуються для інтелектуальної обробки текстових даних з інформаційних потоків та для побудови й аналізу мережевих моделей. Велика увага дослідженням в цій галузі приділена у роботах таких вітчизняних та зарубіжних вчених, як Глушкова В. М., Томащевського В.М., Ланде Д.В., Широкова В.А, Палагіна О.В., Шаронової Н.В., Висоцької В. А., а також Ердоша П., Рені А., Барабаші А.-Л., Альберт Р., Ньюмана Дж., Ваттса Дж. Д., Строґаца С. Г. та інших. Всі вони зробили суттєвий вклад в розвиток теоретичних основ для створення методів і засобів обробки природомовних текстів, а також формування та дослідження складних мереж.

Проте незважаючи на наявні методи та засоби дослідження мережевих структур, в тому числі й отриманих на основі текстових даних, і беручи до уваги стрімкий розвиток інформаційного простору, потреба у розробці нових та удосконалені існуючих лінгвостатистичних методів є актуальною. А отже актуальними є й дослідження, здійснені у цій дистертаційній роботі.

Зв’язок роботи з науковими програмами, планами, темами. Дисертаційну роботу «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу» (державний реєстраційний номер 0117U001454) виконано у відділі спеціалізованих засобів моделювання Інституту проблем реєстрації інформації НАН України відповідно до плану фундаментально-прикладних наукових досліджень, що увійшли до науково-дослідних робіт: «Розробка методів і моделей підтримки прийняття рішень при розпізнаванні інформаційних операцій» (2019-2020 рр., державний реєстраційний номер: 0119U001867), «Розробити механізми підвищення живучості для забезпечення функціональної стійкості систем організаційного управління об’єктів критичних інфраструктур» (2017-2021 рр., державний реєстраційний номер 0117U004106). Також результати роботи та практичні напрацювання були задіяні у рамках ННЦ "Світовий центр даних з геоінформатики та сталого розвитку" Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорького» під час науково-технічної роботи за державним замовленням на науково-технічні (експериментальні) розробки та науково-технічну продукцію «Створення інтегрованої платформи для ситуаційного аналізу соціально-економічних і безпекових явищ» (2021-2022 рр., державний реєстраційний номер: 0121U113470), у науково-дослідній роботі «Створення інформаційно-аналітичного ситуаційного центру для сценарного моделювання кризових і безпекових явищ та вивчення їх впливу на економіку і суспільство» (2021-2022 рр., нак. кер. д.т.н., проф. Ю.П.Зайченко, державний реєстраційний номер: 0121U109764), «Розробка методології та програмно-технічного комплексу для системної оцінки безпекового рівня територій України на основі супутникових даних за умов множинних військових загроз» (2023-2024 рр., державний реєстраційний номер: 0123U102015) та «Розробка програмно-технічного комплексу інтелектуального аналізу неструктурованих даних методами штучного інтелекту та OSINT для планування військових операцій» (2024-2026 рр., державний реєстраційний номер: 0124U000838). Також була здійснена реєстрація авторського права на твір № с202204275 від 19.09.2022 – Комп'ютерна програма

автоматичної побудови мереж термінів на основі аналізу текстових потоків «TermsNet».

Мета і задачі дослідження. Метою дисертаційної роботи є розробити нові методи побудови мережевих моделей предметних галузей на основі текстових корпусів і лінгвостатистичного аналізу текстів та розробити нові методи аналізу побудованих мереж для того, щоб приймати ефективні рішення у відповідних предметних галузях, з якими змістовно пов’язані тексти.

Для досягнення поставленої мети в цій дисертаційній роботі вирішуються наступні задачі:

- здійснити огляд та проаналізувати сучасні лінгвостатистичні методи та підходи, що застосовуються з метою структуризації текстових даних шляхом формування мережевих моделей предметних галузей;
- здійснити огляд наявних методів та алгоритмів, що застосовуються для аналізу й дослідження мережевих структур;
- здійснити огляд та застосувати основні підходи до комп'ютеризованої обробки та аналізу текстових документів;
- запропонувати лінгвостатистичний метод автоматичного екстрагування і виявлення взаємозв’язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів;
- запропонувати форму візуального відображення інформаційного потоку в розрізі фразеологізмів;
- розробити правила визначення напрямків зв’язків між вузлами ненаправленої мережі, сформованої зі слів та словосполучень текстового корпусу, що відноситься до певної предметної галузі;
- розробити новий метод встановлення напрямків зв’язків із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);

- розробити новий підхід до визначення вагових значень зв’язків у мережі термінів;
- розробити нову цілісну методику побудови направлених зважених мереж термінів, як семантичних моделей предметних галузей на основі текстових корпусів;
- запропонувати цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусів;
- розробити новий метод аналізу мереж термінів;
- запропонувати нову методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж.

Об’єктом дослідження є процес структуризації текстових інформаційних потоків, розподілених у вебмережі, у вигляді мережевих моделей.

Предметом дослідження є лінгвостатистичні методи побудови та аналізу мережевих моделей предметних галузей на основі текстових корпусів.

Методи дослідження. В процесі вирішення поставлених задач були використані наступні наукові методи: методи автоматичної обробки та аналізу природної мови та методи комп’ютерної лінгвістики, завдяки яким проводилась попередня комп'ютеризована обробка природномовних текстів, лексичний аналіз та виявлення семантичних зв'язків; методи статистичного аналізу, що застосовувались для виокремлення ключових термінів (слів та словосполучень) із текстових даних; та методи дискретної математики, зокрема, методи теорії графів та складних мереж, завдяки яким здійснювалось побудова мережевих моделей предметних галузей та подальше дослідження й аналіз отриманих моделей.

Наукова новизна одержаних результатів. Наукова новизна роботи полягає у розв’язанні наукової задачі формування мережевих моделей предметних галузей на основі текстових корпусів з використанням лінгвостатистичних методів та подальшого аналізу отриманих мереж термінів.

Під час вирішення поставленої задачі було отримано наступні результати, що містять елементи наукової новизни:

1. запропоновано та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency, або українською – глобальна частота терміна), який на відміну від звичайного статистичного показника TF-IDF дозволяє більш ефективно знаходити ключові та інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми;
2. вперше запропоновано метод виокремлення ключових термінів із текстового корпусу зі застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);
3. вперше запропоновано лінгвостатистичний метод автоматичного екстрагування і виявлення взаємозв'язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів;
4. вперше запропоновано форму візуального відображення інформаційного потоку в розрізі фразеологізмів – Ph-Di діаграму;
5. запропоновано правила визначення напрямків зв'язків між вузлами ненаправленої мережі, сформованої зі ключових слів та словосполучень тематичного текстового масиву, що змістовно відноситься до певної предметної галузі;
6. вперше запропонований та розроблений метод визначення напрямків зв'язків із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);
7. запропоновано та розроблено новий підхід до визначення вагових значень зв'язків у мережі термінів;
8. вперше запропоновано цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусів;
9. вперше запропоновано методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій;

10. вперше запропоновано методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж, та на основі цієї методики запропонована модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Практичне значення одержаних результатів полягає в тому, що їх можна використовувати для структуризації текстових даних, що відносяться до певної предметної галузі. Формування мережевих моделей предметних галузей на основі текстових корпусів допомагає систематизувати, візуалізувати та аналізувати інформацію для різноманітних цілей, від розуміння семантичної структури текстів до підтримки прийняття рішень:

1. Мережі термінів дозволяють розуміти концептуальні зв'язки між термінами чи поняттями в предметній галузі. Це може допомогти уточнити семантичну структуру понять та їх взаємозв'язки.
2. Мережеві моделі допомагають покращити аналіз текстових даних та отримати нові інсайти. Це важливо для розвитку систем штучного інтелекту, які здатні до автоматизованого аналізу та виведення корисної інформації.
3. Створені мережеві моделі можуть бути використані для підтримки роботи з текстами у різних галузях для покращення роботи з текстовою інформацією, сприяючи кращому розумінню та аналізу.
4. Представлена мережева методика порівняння текстових документів може бути використана для виявлення структурних і термінологічних розбіжностей у правовій сфері, що сприятиме парламентському контролю та гармонізації міжнародного права.
5. Мережі термінів можуть бути використані для розробки покращених систем інформаційного пошуку та рейтингування джерел, зокрема, семантичного інформаційного пошуку та систем рекомендацій, оскільки дозволяють покращити точність результатів пошуку шляхом кращого розуміння змісту та семантики запитів користувачів. Це полегшує користувачам знаходження

потрібної інформації та розуміння взаємозв'язків між поняттями, що є критичним у великих обсягах даних.

6. Мережеві моделі текстів можуть бути важливим інструментом для створення, підтримки та управління базами знань у різних організаціях, допомагаючи відфільтрувати та структурувати великі обсяги інформації.
7. Формування мережевих моделей предметних галузей на основі текстових корпусів та подальший їх аналіз дозволяє робити конструктивні висновки та може служити основою для рекомендацій та прийняття ефективних рішень у різних галузях, від бізнесу до академічних досліджень, допомагаючи зрозуміти важливі аспекти та зв'язки між ними.
8. Отримані результати досліджень можуть сприяти подальшому розвитку технологій обробки природної мови та систем автоматичної обробки текстів, полегшуючи автоматизацію обробки та аналізу текстової інформації.

Отже, мережеві моделі предметних галузей на основі текстових корпусів не лише допомагають у структуризації та аналізі інформаційних потоків, але й мають широкий спектр застосувань у багатьох сферах, що робить їх важливим інструментом для подальших досліджень та практичного застосування.

Результати дисертаційної роботи Дмитренка Олега Олександровича на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу» було використано і впроваджено при реалізації програмно-технічних засобів в середовищі Інформаційно-аналітичного ситуаційного центру К І ім. Ігоря Сікорського для комплексного використання та інтелектуального аналізу великих масивів неструктурованих даних різної природи, в тому числі результатів обробки текстів природної мови, в ході виконання низки держбюджетних і договірних науково-дослідних робіт (НДР) та проектів ННЦ «СЦД-Україна», серед яких:

– НДР «Створення інтегрованої платформи для ситуаційного аналізу соціально-економічних і безпекових явищ» (0121U113470), в рамках якої було впроваджено нові методи лінгвостатистичного аналізу надвеликих

масивів текстових даних для оцінювання ставлення суспільства до дій влади на основі аналізу даних з відкритих Інтернет-видань і соціальних мереж.

– НДР «Створення інформаційно-аналітичного ситуаційного центру для сценарного моделювання кризових і безпекових явищ та вивчення їх впливу на економіку і суспільство» (0121U109764), в рамках якої було впроваджено метод формування мережевих моделей предметних галузей (семантичних мереж) на основі текстових корпусів.
– НДР «Розробка методології та програмно-технічного комплексу для системної оцінки безпекового рівня територій України на основі супутникових даних за умов множинних військових загроз» (0123U102015), в рамках якої впроваджено метод виокремлення ключових термінів із текстів із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging).
– НДР «Розробка програмно-технічного комплексу інтелектуального аналізу неструктурованих даних методами штучного інтелекту та OSINT для планування військових операцій» (0124U000838), в рамках якої впроваджено методику порівняння текстових документів (новинних повідомлень в середовищі електронних медіа-ресурсів), що базується на побудові та порівнянні відповідних їм семантичних мереж, та на основі цієї методики впроваджено модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Окрім того, впровадження нових методів лінгвостатистичного аналізу надвеликих масивів текстових даних та нового методу виокремлення ключових термінів із текстових (новинних) повідомлень для формування їх семантичних мереж й вимірювання семантичної близькості відповідних повідомлень в середовищі електронних медіа-ресурсів од інформаційної системи аналітичної обробки інформації Інформаційно-аналітичного ситуаційного центру КПІ ім. Ігоря Сікорського дозволило покращити інтелектуальну обробку та аналіз текстових даних з веб-сайтів та соціальних мереж під час формування дайджестів та отримання нових інсайтів, а реалізація моделі ранжування як окремих документів,

так і джерел інформації у інформаційній системі збору та аналітичного оброблення дозволила вцілому покращити рейтингування джерел на 12%. Також вищезгадані впровадження допомогли підвищити повноту охоплення інформації на 25% за рахунок врахування ширшого переліку інтернет-джерел та медіа-ресурсів, зокрема таких, як канали Telegram і Youtube, а також підвищити загальну швидкість обробки текстових даних та оперативність надання релевантної інформації замовникам у відповідь на їх запит під час інформаційного пошуку.

Особистий внесок здобувача. Дисертаційна робота є результатом самостійного дослідження. У роботах, що виконані у співавторстві, здобувачеві належить: [1] – дослідження статистичних показників важливості термінів, [2, 3] – дослідження нового статистичного показника важливості термінів у тексті GTF, [4, 5, 21, 22] – побудова термінологічних онтологій, як мереж із ключових термінів, [7, 8] – розробка методу встановлення напрямків зв'язків у мережі термінів, [9] – розробка методу визначення вагових значень зв'язків у мережі термінів, [10-12, 16] – розробка методу побудови направлених зважених мереж термінів на основі текстових корпусів, [13, 14, 18] – застосування методу побудови направлених зважених мереж термінів для підтримки прийняття рішень під час розпізнавання інформаційних операцій, [17, 20] – розробка методики виокремлення ключових слів і словосполучень та побудова направлених зважених мереж термінів із застосуванням Part-of-Speech tagging, [23, 25, 28, 31] – застосування методики виокремлення ключових слів і словосполучень та побудова направлених зважених мереж термінів із застосуванням Part-of-Speech tagging, [24, 26, 27, 28] – розробка та застосування методу визначення ступеня подібності текстів з використанням направлених зважених мереж термінів, [29] – дослідження наборів ключових слів-маркерів, які використовуються для автоматизованого розпізнавання мови мовлення, [32] – розробка моделі семантичного пошуку на основі порівняння текстових документів та визначення ступеня подібності (розбіжності) між ними, що базується на побудові та порівнянні відповідних їм семантичних мереж.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи були оприлюднені й обговорювались на таких конференціях:

- XVIII Міжнародна науково-практична конференція «Інформаційні технології і безпека» (ІТБ-2018) (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 27 листопада 2018 р.).
- XVII Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (Фізико-технічний інститут, НТУУ “КПІ імені Ігоря Сікорського”, м. Київ, Україна, 25 - 26 квітня 2019 р.).
- Щорічна науково-технічна конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 15-16 травня 2019 р.).
- XIX Міжнародна науково-практична конференція «Інформаційні технології і безпека» (ІТБ-2019) (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 28 листопада 2019 р.).
- IV Міжнародна конференція «Комп’ютерна лінгвістика та інтелектуальні системи» (CoLInS 2020) (кафедра інформаційних систем та мереж Інституту комп’ютерних наук та інформаційних технологій Львівської політехніки, м. Львів, Україна, 23–24 квітня 2020 р.).
- XVIII Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених «Теоретичні і прикладні проблеми фізики, математики та інформатики» (Фізико-технічний інститут, НТУУ “КПІ імені Ігоря Сікорського”, м. Київ, Україна, 12 – 13 травня 2020 р.).
- XI Всеукраїнська науково-практична конференція «Актуальні проблеми управління інформаційною безпекою держави» (Національна академія Служби безпеки України, м. Київ, Україна, 15 травня 2020 р.).
- XV Міжнародна науково-практична конференція «Математичне та імітаційне моделювання систем. МОДС’2020» (MODS2020) (Чернігівський національний технологічний університет, м. Чернігів, Україна, 29 червня – 1 липня 2020 р.).

- Щорічна науково-технічна конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 28-29 вересня 2020 року)
- 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (м. Київ, Україна, 05–09 жовтня 2020 р.).
- XX Міжнародна науково-практична конференція «Інформаційні технології і безпека» (ІТБ-2020) (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 10 грудня 2020 р.).
- XII Всеукраїнська науково-практична конференція «Актуальні проблеми управління інформаційною безпекою держави» (Національна академія Служби безпеки України, м. Київ, Україна, 26 березня 2021 року).
- Перша науково-практична конференція «Парламентський контроль в умовах децентралізації державної влади та цифрової трансформації в Україні: стан та проблеми» (м. Київ, Україна, 30 березня 2021 року).
- V Міжнародна конференція «Комп'ютерна лінгвістика та інтелектуальні системи» (CoLInS 2021) (Кафедра інтелектуальних комп'ютерних систем НТУ ХПІ", м. Харків, Україна, 22-23 квітня 2021 року).
- Міжнародна наукова-технічна конференція «Інтелектуальні технології лінгвістичного аналізу» (Факультет кібербезпеки, комп'ютерної та програмної інженерії Національного авіаційного університету, м. Київ, Україна, 18-19 жовтня 2021р.).
- XX Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (Фізико-технічний інститут, НТУУ "КПІ імені Ігоря Сікорського", м. Київ, Україна, 15 червня 2022 р.).
- Щорічна підсумкова науково-технічна конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 27-28 вересня 2022 р.).
- 2022 IEEE 3nd International Conference on System Analysis & Intelligent Computing (SAIC) (м. Київ, Україна, 04–07 жовтня 2022 р.).

– XXIII Міжнародна науково-практична конференція «Інформаційні технології і безпека» (ІТБ-2023) (Інститут проблем реєстрації інформації НАН України, м. Київ, Україна, 30 листопада 2023 року).

Публікації. За результатами дисертаційних досліджень опубліковано 34 наукові праці (в тому числі 5 – одноосібні [6, 15, 19, 30, 32]). Серед них 8 наукових статей опубліковані в фахових наукових виданнях України, серед яких за спеціальністю 122 «Комп'ютерні науки» – 6 статей [1, 9, 18, 19, 28, 32], не за спеціальністю – 2 [4, 26], та 1 стаття [29] опублікована у фаховому закордонному журналі, що належить до квартилю Q3 за спеціальністю здобувача. За матеріалами виступів на 19-ти науково-технічних конференціях опубліковано 25 робіт, серед них 9 тез доповідей наукових конференцій [6, 12, 13, 15, 21, 22, 24, 30, 34], 6 статтей конференцій [2, 5, 7, 11, 17, 27] та 5 статтей [3, 8, 10, 20, 23], що розміщені в міжнародному електронному виданні CEUR Workshop Proceedings, що індексується базою Scopus. Розширені та доопрацьовані матеріали конференцій [16, 31] та роботи [14, 25, 33], які увійшли як окремі розділи до серій книг за спеціальністю 122 «Комп'ютерні науки», також індексуються Scopus та WoS. Також було оформлено 1 свідоцтво про реєстрацію авторського права на твір [35]. Загальна кількість публікацій у наукових виданнях, включених на дату опублікування до переліку наукових фахових видань України за спеціальністю 122 «Комп'ютерні науки» та у періодичних наукових виданнях, проіндексованих у базах даних Web of Science Core Collection та/або Scopus, з урахуванням числа співавторів та першого-третього квартилів (Q1-Q3) відповідно до класифікації SCImago Journal and Country Rank або Journal Citation Reports, становить 13 наукових публікацій.

Структура дисертації. Дисертація складається зі вступу та чотирьох розділів. Загальний обсяг дисертаційної роботи становить 170 сторінок, серед яких основну частину складають 131 сторінка. Також робота містить 24 рисунки та 13 таблиць.

РОЗДІЛ 1. ОГЛЯД ЛІНГВОСТАТИСТИЧНИХ МЕТОДІВ ФОРМКВАННЯ Й АНАЛІЗУ МЕРЕЖЕВИХ МОДЕЛЕЙ

У цьому розділі проводиться огляд стану проблеми та наукових розробок, яким присвячена тема дисертації. Зокрема було здійснено огляд сучасних комп'ютерно-лінгвістичних підходів та методів автоматичного аналізу текстових інформаційних потоків з метою виявлення знань з предметної галузі з якою змістовно пов'язані текстові дані. Було встановлено, що існує декілька підходів, зокрема такі як статистичний та лінгвістичний. В основі статистичного підходу лежить припущення, що зміст тексту відображається за допомогою найбільш уживаних у тексті слів. Тож ідея статистичного аналізу полягає у підрахунку кількості входжень або повторень певного слова у текстовому документі та обчисленні його вагового значення. Тож було здійснено огляд існуючих методів статистичного зважування термінів. У випадку лінгвістичного підходу текст розглядається як організована у деякий спосіб послідовність графем та рядків. Лінгвістичний підхід розрізняє наступні рівні обробки текстових даних, які детально описані в цьому розділі, зокрема: графемний, де відбувається обробка послідовності символів та виокремлення структури тексту (окремих розділів, параграфів, речень та слів); морфологічний, де здійснюється визначення морфологічних ознак та характеристик слова та дослідження його можливих форм, і як наслідок присвоєння слову ряду атрибутів, зокрема частини мови; синтаксичний, що пов'язаний з проблемою складності структури речень у флективних мовах та семантичний підхід, що дозволяє перетворювати вихідний текст у певну мережеву модель – «семантичну мережу», «семантичний граф». Розглянуто основні ідеї семантичного пошуку як одного із найперспективніших видів автоматизованого повнотекстового інформаційного пошуку, що сприяє розвитку інформаційно-пошукових систем.

1.1. Огляд стану та проблем сучасного інформаційного простору

Природна мова – це історично сформована усна (звукова) та знакова (графічна) інформаційна система, що утворилася у результаті людської взаємодії задля передавання та закріплення накопиченої у процесах пізнання та спілкування інформації. Знакова природна мова є засобом передання інформації, що описує найрізноманітніші сфери життя та галузі діяльності. Тож її обробка та подальший аналіз може дати цінні знання про ту галузь, з якою змістовно пов’язані текстові дані.

Наразі в умовах стрімкої цифровізації суспільства особливо актуальним напрямком сучасної науки є дослідження та вивчення комп'ютерної репрезентації і аналізу природної мови. Адже постійне хоча й далеко не завжди регулярне оновлення інформації спричинило попит на систематичне відстеження тенденцій і суспільних явищ, процесів та навіть думок у динамічному та постійно оновлюваному інформаційному просторі.

Загалом, під інформаційним простором прийнято розуміти сукупність інформаційних ресурсів та технологій, завдяки яким вони використовуються та супроводжуються, а також інформаційних і телекомунікаційних систем, які в сукупності утворюють інформаційну інфраструктуру. Стрімкий розвиток інформаційно-комунікаційних технологій та систем і глобалізація інформаційного простору призвели до не менш стрімкого розвитку інформаційних ресурсів у мережі Інтернет [36, 37], а отже – до збільшення динамічних інформаційних потоків, зокрема, текстових, що розподілені в вебмережі [38] . Такий інформаційний сплеск, або так званий інформаційний вибух, характеризується стрімким збільшенням об’ємів даних, які продукуються елементами інформаційного простору, зокрема, документами та найрізноманітнішими джерелами даних – файлами, електронними листами, вебсторінками не залежно від форматів їх подання. Важливим є той факт, що обсяг вищезгаданих даних подвоюється приблизно кожні 18 місяців [39]. Унаслідок цього за п'ять попередніх років людством було продуковано даних більше, ніж за всю попередню історію

[40]. Наприклад, у глобальній мережі Інтернет налічується більше сотні трильйонів документів. За даними сервісу Netcraft [41] станом на серпень 2021 року кількість активних вебсайтів становить понад 1,2 мільярда, а кількість унікальних доменів – більш ніж 200 мільйонів (рис. 1.1).

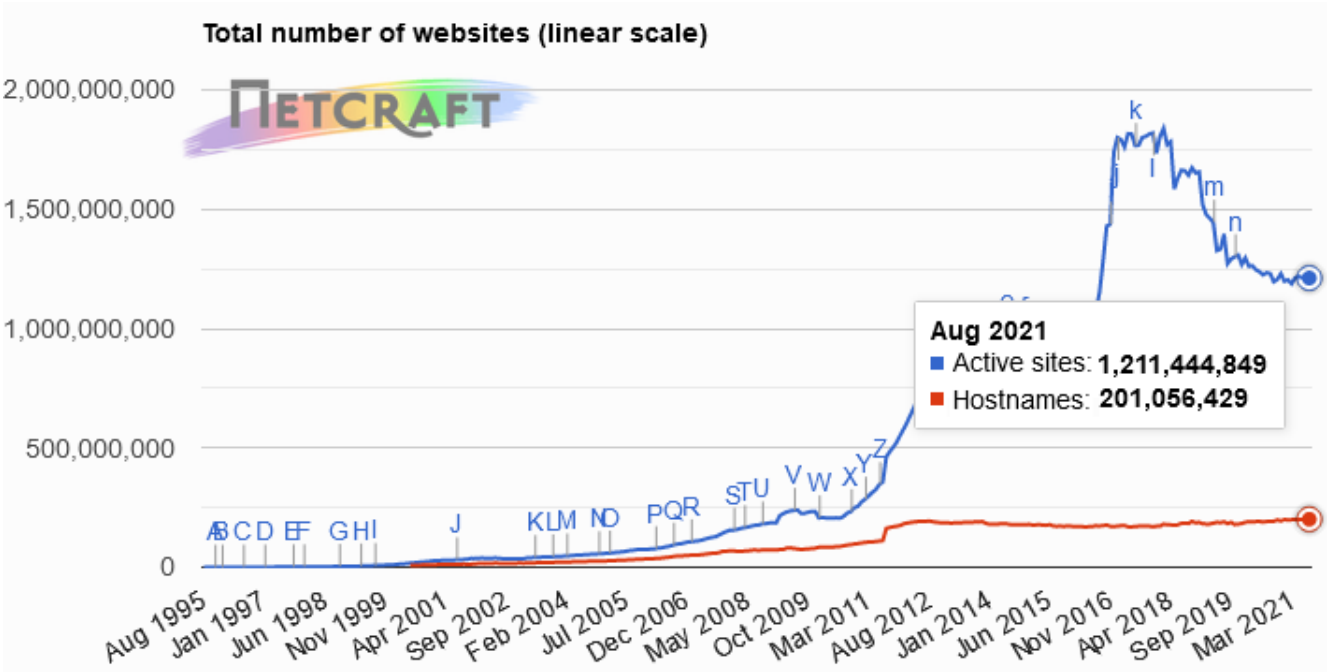

Рис. 1.1. Динаміка кількості сайтів в усіх доменах за даними сервісу Netcraft

У всьому світі Інтернет продовжує змінювати наш стиль взаємодії та обміну інформацією. За даними сайту Statista [42] у 2019 році кількість користувачів Інтернету у всьому світі становила 3,97 мільярдів, що означає, що більше половини населення планети наразі підключено до всесвітньої павутини. Станом на вересень 2021 року кількість інтернет-користувачів складає вже більше 5 мільярдів (рис. 1.2). У 2023 році кількість користувачів Інтернету в усьому світі становила 5,18 мільярда [42], що означає, що близько двох третин населення світу зараз підключено до всесвітньої мережі. Це породжує зростання активності у мережі Інтернет.

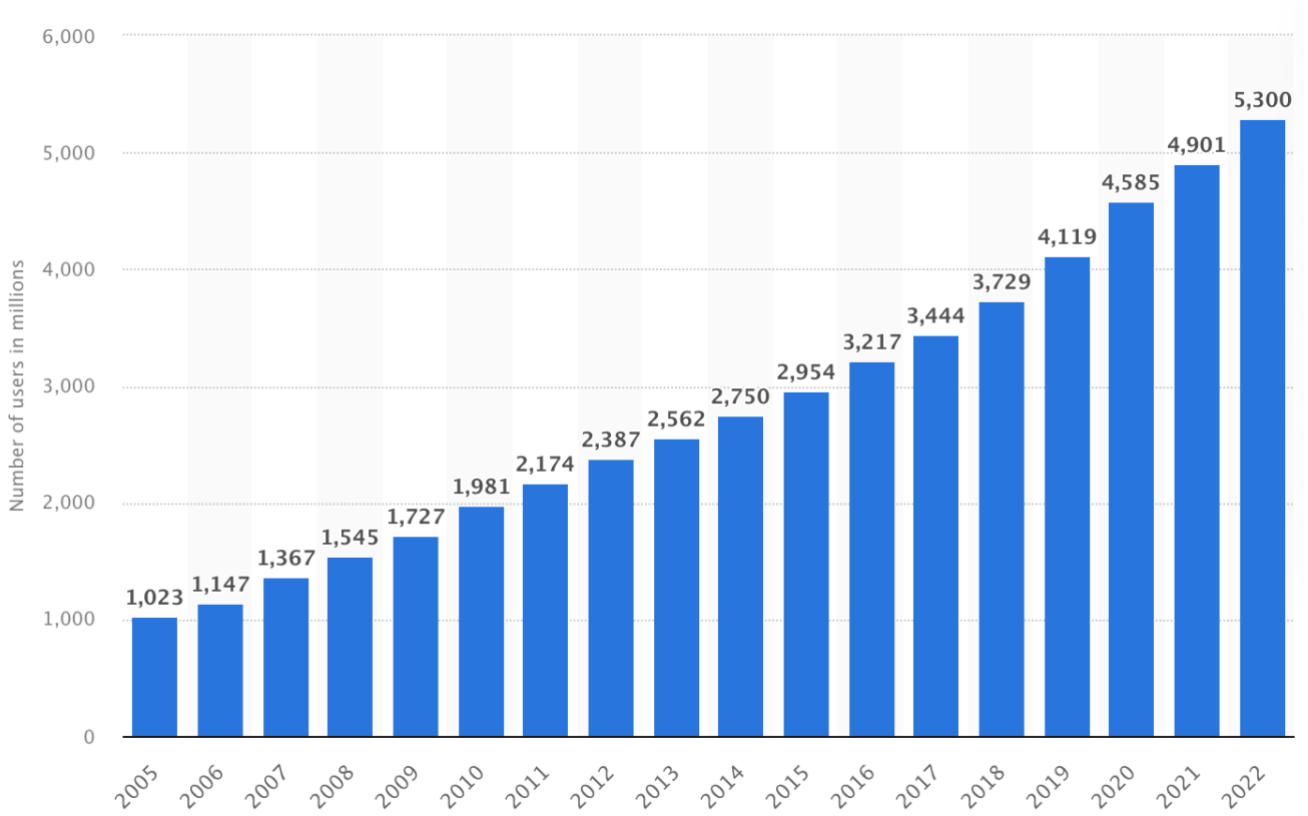

Рис. 1.2. Динаміка кількості інтернет-користувачів згідно з Statista.com [42]

Наприклад, у 2010 році Google зайняв перше місце як найпопулярніший та найбільш відвідуваний вебсайт у світі, і з тих пір утримує цю позицію, збираючи станом на листопад 2022 року більш ніж 88,4 мільярда відвідувань глобальної онлайн-аудиторії на місяць (рис. 1.3).

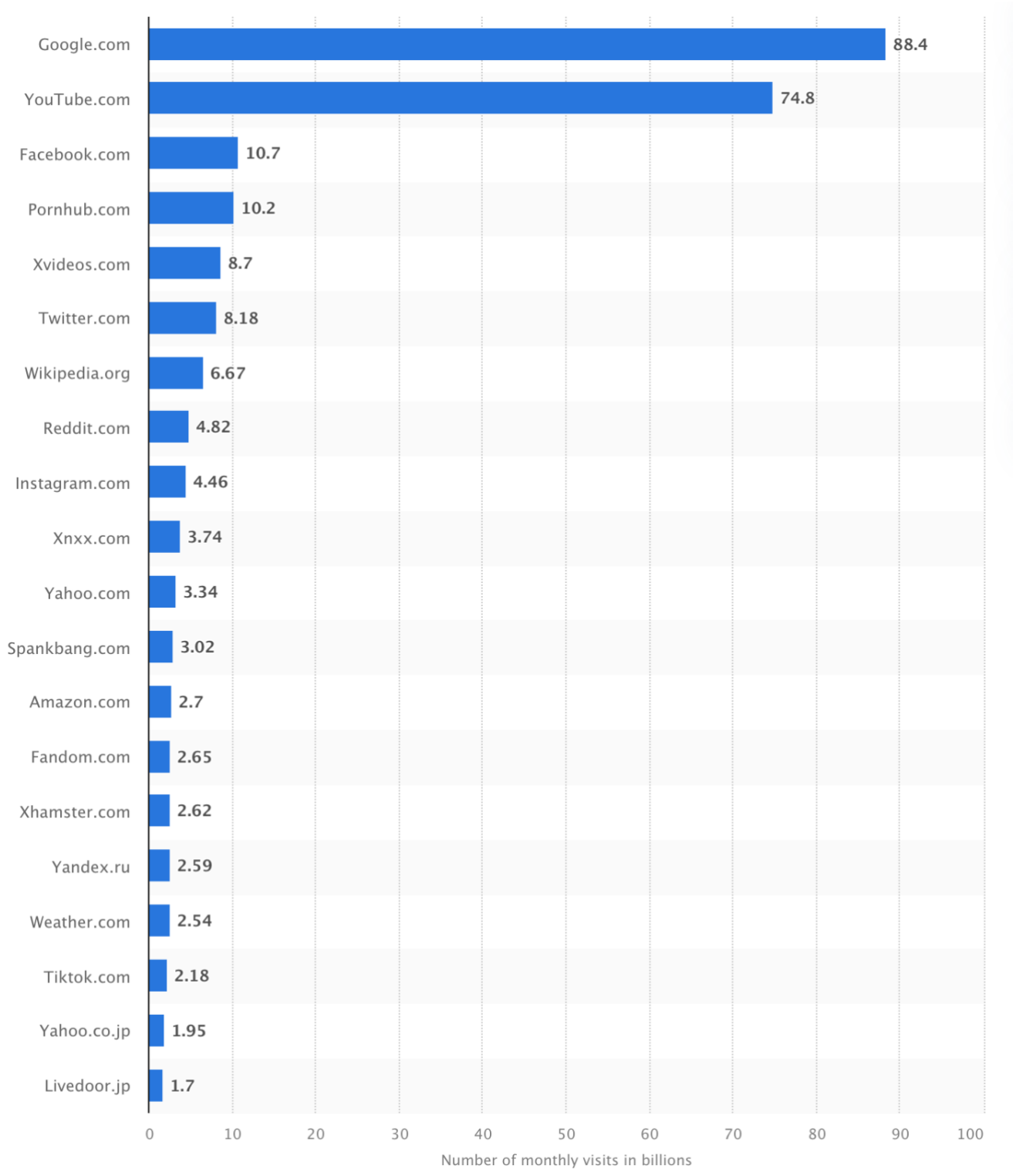

Рис. 1.3. Найпопулярніші вебсайти у світі станом на листопад 2022 року за загальною кількістю відвідувань (у мільярдах) згідно з Statista.com [43]

Також за даними сайту Statista [44] станом на третій квартал 2023 року щомісячно кількість активних користувачів соціальної мережі Facebook становить майже 3 мільярда (активні користувачі – ті, хто увійшов у Facebook протягом

останніх 30 днів), що робить її найбільшою соціальною мережею у світі. Відомо також, що активні користувачі Facebook кожного дня генерують більш як 4 петабайти даних. Протягом останнього кварталу 2021 року компанія заявила, що 3,59 мільярда людей щомісяця користуються принаймні одним із основних її продуктів (Facebook, WhatsApp, Instagram або Messenger).

Інтернет-трафік – це один з найважливіших показників для оцінки продуктивності та успіху в Інтернеті. Він представляє собою обсяг даних, що надіслані та отримані через комп'ютерну мережу за певний період часу. Ще у 2016 році обсяг інтернет-трафіку становив 1 зеттабайт (1 099 511 627 776 гігабайт). За прогнозами, загальний обсяг створених, скопійованих та споживаних даних у всьому світі швидко зростатиме. У порівнянні з 2020 роком, коли загальний обсяг даних у мережі Інтернет сягав 64,2 зеттабайт, протягом наступних п'яти років до 2025 року прогнозується, що створення глобальних даних зросте до більш ніж 180 зеттабайт.

Наприклад, у 2020 році кількість створених та відтворених даних досягла нового максимуму. Зростання було вищим, ніж очікувалося раніше, через зростання попиту через пандемію COVID-19. Це пов’язано з тим, що все більше людей працювали та вчилися вдома, а також частіше користувалися всесвітньою мережею задля проведення свого дозвілля та в пошуках розваг. Так наприклад, у вівторок, 10 березня 2020 року, центри обробки даних у всьому світі зафіксували рекордно високі обсяги інтернет-трафіку. Аналітики пояснюють такий стрімкий ріст активності інтернет-користувачів пандемією коронавірусу та випуском нової гри із серії Call of Duty. Один з найбільш завантажених у світі мережевих вузлів, франкфуртський DE-CIX, зафіксував максимальний рівень трафіку за всю історію, понад 9,1 Тбіт/с. [45] Надзвичайна ситуація, оголошена в ряді країн у зв'язку з пандемією коронавірусу, здійснила істотний вплив на інтернет-послуги та життя користувачів Інтернету, відзначають в американській компанії, що надає мережеві послуги – Cloudflare [46]. У звіті компанії говориться, що світовий інтернет-трафік зріс на величину від 10% до 40%, залежно від регіону. Безперечно такий сплеск породив численну кількість публікацій на різноманітних інтернет-ресурсах.

1.2.Проблеми розвитку інформаційних потоків

Розвиток мережі Інтернет і розширення інформаційного простору спричиняють також ряд специфічних проблем, пов’язаних, в першу чергу, зі стрімким збільшенням об’ємів даних, або так званим інформаційним перевантаженням [47]. Зі збільшенням кількості інформаційних потоків зростає і частка неструктурованих або слабоструктурованих текстових даних, в тому числі й непотрібних та «шумових». Також зростає і кількість дублюючих даних. Все вищезгадане ускладнює пошук й отримання необхідної та релевантної інформації. Наприклад, основна частина (близько 95%) даних є неструктурованими, і лише зовсім мала (близько 5%) – це різні бази даних, де зберігається структурована інформація, яка може бути використана під час прийняття рішень. Тож у зв’язку з розвитком Інтернету і розширенням інформаційного простору перед інформаційним суспільством постає також і ряд специфічних проблем, пов’язаних зокрема з критичною невідповідністю між розвитком сучасних інформаційних систем і збільшенням динамічних текстових інформаційних потоків у глобальних комп’ютерних мережах [38]. А саме, проблема полягає у відсутності підходящих технологічних рішень та у неспроможності наявних систем обробляти величезні об’єми неструктурованих текстових даних й виокремлювати з них знання з тією ж самою швидкістю, з якою відповідні дані продукуються й накопичуються. Тож питання зберігання та подальшої комп’ютерної обробки даних з метою екстрагування знань та подальшої їх структуризації є важливим у сучасному інформаційному середовищі.

Сьогодні існує кілька глобальних проблем, пов'язаних із розвитком інформаційних потоків та інтернет-контенту, серед яких дві найголовніші [48]:

- прогрес у сфері продукування інформації веде до зниження рівня поінформованості людей;
- інтенсивність зростання обсягів шумової інформації в багато разів перевищує інтенсивність зростання обсягів корисної інформації.

Новий рівень розвитку мережного інформаційного простору зумовлює необхідність створення та розвитку адекватних моделей інформаційного простору, інформаційних потоків, мережевого пошуку. У зв'язку з цим виникає інтерес до підходів, що ґрунтуються на розумінні інформації як заходи впорядкованості деякої системи і, відповідно, до статистичних методів її обробки. Для організації ефективної комунікації в мережах сьогодні доводиться постійно повертатися до джерел теорії інформації, понять ентропії, теорії Шеннона, рівнянь Больцмана та ін., що зумовлює широкі перспективи застосування потужного апарату математики та фізики у вирішенні теоретико-інформаційних завдань.

1.3.Основні поняття інформаційного пошуку

Доступ користувачів до сучасних інформаційних мереж, ефективне задоволення їхніх інформаційних потреб можливе лише за допомогою розвинених засобів навігації у цих мережах. Основним інструментом у своїй виступають інформаційно-пошукові системи, які забезпечують пошук у гігантських обсягах текстової інформації. Користувач звертається до інформаційно-пошукової системи (ІПС) з певним запитом, який відображає його інформаційні потреби. Таким запитом зазвичай є текст, поданий природною мовою, яка зрозуміла людині – користувачу ІПС. Завдяки ІПС пошуковий запит – текст – ставиться у відповідність кожному елементу пошукового масиву, який формується та періодично оновлюється ІПС та її адміністраторами. Такий масив заздалегідь вводиться в пошукову систему у вигляді документів поданих зазвичай природною чи близькою до неї мовою, і опісля здійснюється їх індексування [49]. Індексування – процес перетворення тексту у формальну мову інформаційно-пошукової системи – інформаційно-пошукову мову (ІПМ).

Традиційні пошукові системи використовують пошуковий образ документа (ПОД) – текст, представлений інформаційно-пошуковою мовою, який ставиться у однозначну відповідність документу та відображає ключові ознаки цього документу, які необхідні для його пошуку під час запиту.

Загалом процес індексування включає наступні взаємопов'язані етапи:

1. Аналіз змісту документа;

2. Виявлення та відбір змістових компонентів у змісті документа (вибір понять, що характеризують зміст документа);

3. Формування переліку ключових слів, що використовується для створення пошукового образу документа (ПОДа) (ухвалення рішення про склад ПОДа);

4. Нормалізація ключових слів за формою та змістом (оформлення відібраних смислових компонентів як понять у розумінні індексування та відповідно до системи граматичних засобів даної ІПМ);

5. Надмірне індексування;

6. Заповнення поля ПОДа з назвою «Ключові слова».

Існує декілька способів перетворення тексту у формальну внутрішню мову інформаційно-пошукової системи. Одним з таких є координатне індексування [50], коли поданому природною мовою документу присвоюються набір ключових слів, які визначають основний зміст цього документа. Метод координатного індексування базується на положенні, що основний зміст документа та інформаційної потреби може бути з достатнім ступенем точності і повноти виражено відповідним списком так званих ключових слів, які явно або в прихованому вигляді містяться в тексті. Під ключовими словами в даному випадку розуміються найбільш суттєві інформативні слова та словосполучення, що володіють називною (номінативною) функцією. Називні слова виділяють предмет, вказуючи на нього. Наприклад, власні імена є найбільш відомими представниками називних слів. Окрім називних в якості ключових слів можуть виступати також відповідні чисельні характеристики, хронологічні дані, діапазони температур, тисків і т.д. Також ключовими можуть бути лексичні одиниці, що можуть являти собою слова, стійкі словосполучення, абревіатури, символи, дати, загальноприйняті скорочення, а також лексично значущі компоненти складних слів та еквівалентні їм кодові чи символічні позначення штучної мови, наприклад, коди класів класифікаційної системи, які позначають окреме поняття. Виокремлений з тексту документа список ключових слів утворює пошуковий образ. Вибрані з

тексту інформативні слова, які застосовуються як ключові, при необхідності доповнюються, уточнюються, змінюються. Практичний досвід показує, що для координатного індексування одного документа зазвичай достатньо 6-12 ключових слів.

Надмірне індексування здійснюється шляхом включення до пошукового образу документа близьких за змістом лексичних одиниць інформаційно-пошукової мови для підвищення повноти та якості пошуку. Також допускається дублювання дублювання ключових слів для зручності пошуку.

Тож координатне індексування – це спосіб вираження основного змісту документа або інформаційної потреби користувача у вигляді певної сукупності ключових слів. Координатне індексування називають також методом координації понять, корелятивним індексуванням, унітерм-індексуванням, асоціативним індексуванням, комбінаторним індексуванням тощо. Існують два способи координатного індексування: вільне – безпосереднє виявлення та виокремлення із тексту ключових слів у тій формі, у якій вони подані у тексті, без урахування їх словоформ та семантико-синтаксичних взаємозв’язків у тексті; та контрольоване – пошуковий образ документа формується на основі тих слів, які входять до інформаційно-пошукового тезаурусу [51], де враховані всі можливі форми слова, зазначені його морфологічні, синонімічні та асоціативні особливості.

Під час вільного координатного індексування ключові слова в пошукових образах не пов'язані одне з одним і функціонують самостійно. Для відшукання документів, що відповідають певному інформаційному запиту, потрібно виконати деякі логічні операції над класами, які позначені ключовими словами пошукових образів документів. У найпростішому випадку, коли пошуковий запит сформульовано у вигляді логічного добутку (кон’юнкції) певної множини ключових слів, то документ вважається таким, що відповідає на інформаційному запиту і підлягає видачі, якщо в пошуковому образі цього документа одночасно містяться всі ключові слова пошукового запиту. Проте спосіб вільного координатного індексування містить і ряд недоліків, що впливають на забезпечення

високої якості інформаційного пошуку. Серед таких недоліків можна виділити наступні [52]:

1. Хибна координація. Предмет інформаційного запиту – «Інформаційні системи економіки». Пошуковий запит сформульовано так: *інформаційні системи*, *економіка*. У відповідь на такий пошуковий припис документальна ІПС видасть як релевантний запиту документ «Інформаційна система з економічного та соціального планування», так і документ «Економіка інформаційних систем», що не відповідає запиту, оскільки обидва пошукові образи містять ключові слова *інформаційні системи* та *економіка*. У цьому випадку недостатньо використовувати у запиті лише координатний зв'язок між ключовими словами у пошуковому образі документа.
2. Неповна координація. Предмет інформаційного запиту – «Внесок постачальника в розробку електронних каталогів». У цьому випадку пошуковий запит може бути сформульований так: *постачальники*, *електронні каталоги*, *технологія*. У результаті пошуку на такий запит документальна ІПС поверне нерелевантний документ «Внесок користувача у розробку електронних каталогів: думка постачальника», оскільки пошуковий образ документа містить ключові слова *постачальники*, *користувачі*, *електронні каталоги*, *розробка*. Причина видачі нерелевантного документа полягає у тому, що з формулювання пошукового запиту було використано ключові слова, які достатні для предмета інформаційного запиту, але недостатні для предмета документа.
3. Синонімія та полісемія ключових слів. Припустимо, що предмет інформаційного запиту – «Застосування анкетування щодо інформаційних запитів користувачів». Пошуковий запит сформульовано так: *інформаційна потреба*, *користувачі*, *анкетування*. Документальна ІПС при цьому не поверне у відповідь на такий запит явно релевантний документ «Вивчення інформаційних потреб

користувачів Канадської інформаційної системи з полярних досліджень на основі анкетного опитування», під час індексування якого були використані ключові слова *інформаційні системи*, *потреби*, *користувачі*, *анкетне опитування*. Причина невидачі релевантного документа полягає в тому, що замість ключового слова *анкетування* у пошуковому образі документа використано його синонім *анкетне опитування*.

4. Невизначеність родо-видових зв'язків між ключовими словами. Нехай предмет інформаційного запиту – «Бібліотечна справа у Європі». Тоді пошуковий запит буде сформульовано так: *бібліотечна справа*, *Європа*. У відповідь на такий пошуковий запит не буде, наприклад, видано відповідний інформаційному запиту документ «Школи, бібліотеки та нова політична система Угорщини», оскільки в його пошуковому образі є ключові слова *бібліотечна справа* та *Угорщина*.
5. Хибні синтагматичні зв'язки. Нехай предмет інформаційного запиту – «Передача електроенергії із Шотландії до Англії», якому відповідатиме пошуковий запит: *передача*, *електроенергія*, *Шотландія*, *Англія*. Документальна ІПС видає документ «Передача електроенергії з Англії до Шотландії», хоча цей документ не відповідає інформаційному запиту. В даному випадку ІПС видала нерелевантний документ, хоча мав місце точний збіг пошукового запиту з ПОД. Звідси випливає, що для запобігання видачі нерелевантних документів необхідно, щоб ключові слова ПОДа та ключові слова пошукових запитів можна було пов'язувати сильнішими синтагматичними зв'язками, ніж зв'язками звичайної координації.

Наведені приклади показують, що для суттєвого підвищення якості інформаційного пошуку, що базується на застосуванні координатного індексування, необхідно:

1) усунення синонімії та полісемії ключових слів, що використовуються як лексичні одиниці ІПМ;

2) побудова спеціальних словників, таблиць чи схем, у яких були б виражені найбільш істотні парадигматичні зв’язки між ключовими словами;

3) розробка для дескрипторної ІПМ такого синтаксису, який дозволяв би використовувати при побудові ПОД та пошукових запитів не тільки просту координацію дескрипторів, а й сильніші синтагматичні зв'язки.

Перша задача відноситься до галузі семантики, тобто до аспекту відношень слів до предметів та явищ, які вони позначають, а друге завдання – до галузі відношень між предметами та явищами, що позначаються словами. Сукупність методів та засобів, що застосовуються для вирішення цих двох завдань, називається контролем за словниковим складом ІПМ, а індексування – контрольованим індексуванням. Завдяки введенню такого контролю забезпечується використання однакових ключових слів для індексування однакових за змістом документів та інформаційних запитів.

Контроль за ключовими словами може мати різні ступені. За відсутності контролю для координатного індексування документа або інформаційного запиту ключові слова вибираються безпосередньо з тексту документа без урахування того, які ці ключові слова вже використовувалися раніше для індексування таких або близьких за змістом документів та інформаційних запитів. В цьому випадку не усувається синонімія, полісемія та омонімія ключових слів, а їх граматичні форми навіть не приводяться до нормального вигляду. Індексування інформаційних запитів у цьому випадку має проводитися дуже ретельно і з надмірністю, яка потрібна для нейтралізації негативних явищ, які породжуються відсутністю словникового контролю під час координатного індексування документів.

При повному контролі за словниковим складом ІПМ дозволено використовувати для індексування документів та інформаційних запитів лише дескриптори, тобто такі ключові слова, які містяться в певному нормативному списку (наприклад, тезаурус). Тезаурус являє собою словник зі спеціально-організованих лексичних одиниць ІПМ (дескрипторів) та лексичних одиниць природної мови. Дескриптори – це призначені для координатного індексування документів та інформаційних запитів нормативні ключові слова, які за певними

правилами відібрано з основного словникового складу тієї чи іншої природної мови та у яких штучно (за допомогою відповідних посилань та позначень) усунуто синонімію, полісемію та омонімію. Для групи ключових слів текстових документів певної предметної області, які мають однакове або подібне значення (синонімічні слова та словосполучення) ставиться у однозначну відповідність один дескриптор, а багатозначному слову може відповідати декілька дескрипторів. При цьому у такому списку або словнику враховано семантичні зв'язки між словами у тексті. Тезаурус і граматика складають ІПМ.

Дескрипторною мовою називається спеціальна ІПМ, словниковий склад якої складається з дескрипторів, а граматика зі способу побудови ПОД та пошукових запитів шляхом координації відповідних дескрипторів.

Тож традиційні інформаційно-пошукові системи поділяються на системи з вільним та контрольованим словниковим складом. Під час вільного координаційного індексування словник поповнюється автоматично після появи нового документа чи набору документів – відбувається актуалізація словника. Відповідно, це передбачає також оновлення образів документів, тобто їх переіндексацію. Така процедура вимагає попередньої фіксації поточного словника й подальшого оновлення та актуалізації даних у базі даних – повне перезавантаження бази даних. Після відновлення словника відбувається перезавантаження документів – їх переіндексація у відповідності з новим словником. Оскільки процедура актуалізації даних є ресурсозатратною, то для її виконання вимагається зупинка системи на час оновлення.

Оскільки одні сторінки та вебресерси можуть зникати, а інші – додаватися, то для відслідковування таких змін розробляються спеціальні роботи, які проводять сканування мережі. Завдяки такому роботу система інформаційного пошуку знаходить нові вебсторінки, формує для них пошуковий образ, приписуючи відповідні ключові слова та проводить індексування цих вебсторінок. Для індексування тектової інформації та формування ПОД найкраще підходять вебдокументи формату HTML.

Отже, якість та швидкість роботи ІПС залежить, в першу чергу, від вибраного підходу індексування.

1.4.Ефективність інформаційного-пошуку

Інформаційно-пошукова система повинна видавати документи, що релевантні пошуковому запиту. Однак, як було зазначено у попередньому параграфі, на практиці спостерігається, що деякі документи, які насправді є релевантними, не видаються пошуковою системою, а замість них видаються зайві, які не відповідають – так званий «шум». Тож загалом масив документів, знайдених інформаційною-пошуковою системою, можна розділити за двома критеріями: на видані та невидані – за одним критерієм, і релевантні та нерелевантні – за іншим. Існує багато характеристик пошуку, серед яких основними визнані дві – це повнота (recall) і точність (precision). Також велика увага нині приділяється такій характеристиці, як пертинентність. Ця характеристика інформаційно-пошукових систем означає відповідність отриманих у результаті пошуку документів інформаційним потребам користувача, а не формальній відповідності документа запиту. Для обчислення показників якості пошуку прийнято розглядати таблицю, яку заповнюють за результатами пошуку навчальної колекції документів.

Таблиця 1.1.

Таблиця результатів пошуку

Документи	Видані	Невидані
Релевантні	a	c
Нерелевантні	b	d

Завдяки цій таблиці розраховуються такі показники інформаційного пошуку, як повнота, точність, якість, помилка, F-міра та середня точність.

Коефіцієнт повноти (англ. recall) виражається наступним чином:

$$\boldsymbol{recall} = \frac{\boldsymbol{a}}{\boldsymbol{a} + \boldsymbol{c}} \tag{1.1}$$

Коефіцієнт точності (англ. precision) розраховується за допомогою наступної формули:

$$\boldsymbol{precision} = \frac{\boldsymbol{a}}{\boldsymbol{a+b}} \tag{1.2}$$

Якість (англ. accuracy) інформаційного пошуку обчислюється наступним відношенням:

$$\boldsymbol{accuracy} = \frac{\boldsymbol{a+d}}{\boldsymbol{a+b+c+d}} \tag{1.3}$$

Помилка (англ. error) інформаційного пошуку виражається як:

$$\boldsymbol{error} = \frac{\boldsymbol{a+d}}{\boldsymbol{a+b+c+d}} \tag{1.4}$$

F-міра (англ. F-measure):

$$\boldsymbol{F} = \frac{\boldsymbol{2}}{\frac{\boldsymbol{1}}{\boldsymbol{recall}} + \frac{\boldsymbol{1}}{\boldsymbol{precision}}} \tag{1.5}$$

Середня точність (англ. average precision):

Додатково до розглянутих пошукових характеристик інформаційно-пошукових систем велике значення мають такі технологічні характеристики, як:

- швидкість обробки запитів;
- повнота охоплення ресурсів;
- доступність, тобто ймовірність отримання відповіді системи;
- знаходження документів, подібних до знайдених;
- можливість уточнення запитів;
- можливість підключення перекладачів тощо.

Повнота охоплення ресурсів Інтернет є одним із двох головних аспектів характеристики повноти мережевої інформаційно-пошукової системи. Наступний аспект пов'язаний з повнотою інформації, яка видається користувачеві на його запит.

Також результат інформаційно-пошукової системи повинен задовольняти умову релевантності. Під релевантністю розуміється доречність знайдених результатів відповідно до пошукового запиту. Також на практиці використовується ще одне поняття – пертинентність. Пертинентність означає відповідність знайдених

результатів потребам користувача (відношення корисної інформації до всього обсягу отриманої).

1.5.Огляд моделей традиційного інформаційного пошуку

Традиційний інформаційний пошук включає в себе використання різних моделей та методів знаходження інформації [53]. На сьогодні відомі такі моделі інформаційного пошуку, як теоретико-множинні (булева модель, розширена булева модель, модель нечітких множин), алгебраїчні (векторна, латентно-семантична, нейромережева) та ймовірнісні які застосовуються в інформаційно-пошукових системах (рис. 1.4).

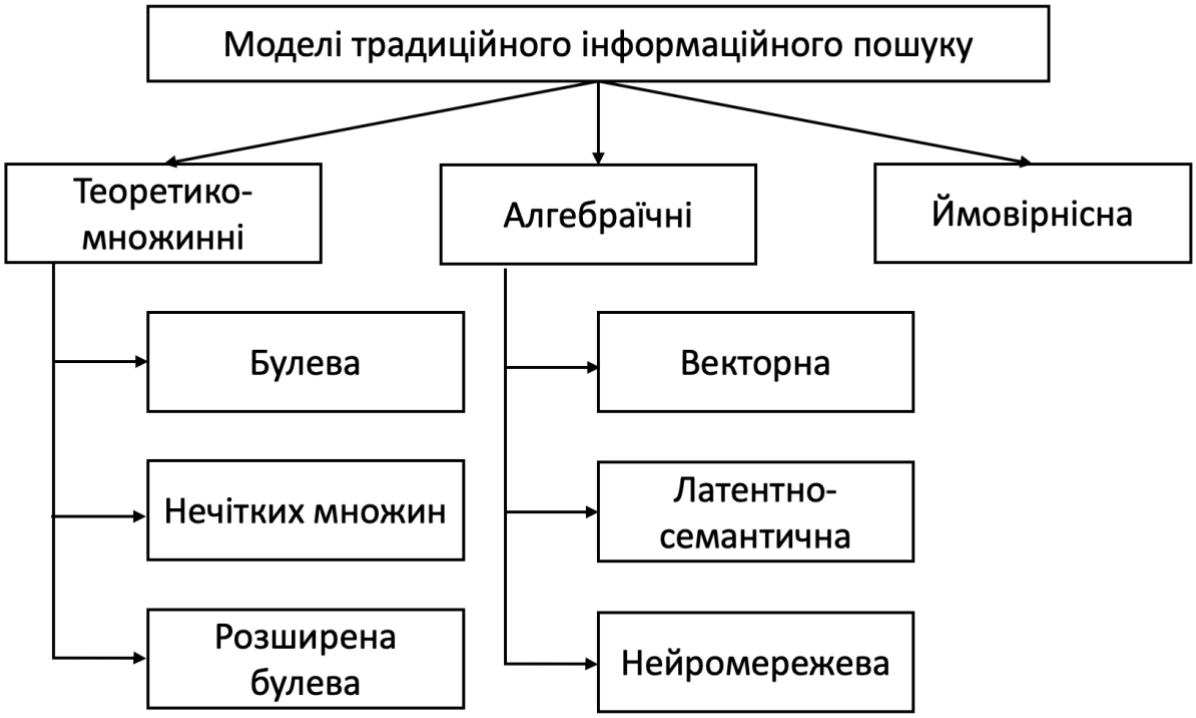

Рис. 1.4. Моделі моделей традиційного інформаційного пошуку

Булева модель

Булева модель інформаційного пошуку є однією з основних моделей, яка використовується для пошуку інформації [54]. Вона базується на принципах

булевої логіки і використовує логічні оператори AND, OR, NOT для формування пошукових запитів.

У булевій моделі, документи та запити представляються у вигляді множин слів або термів. Запити формуються за допомогою логічних операторів:

- AND: повертає документи, які містять всі терміни з запиту. Наприклад, запит "кіт AND собака" поверне документи, які містять як "кіт", так і "собака".
- OR: повертає документи, які містять хоча б один термін з запиту. Наприклад, запит "кіт OR собака" поверне документи, які містять або "кіт", або "собака".
- NOT: виключає документи, які містять вказаний термін з запиту. Наприклад, запит "кіт NOT собака" поверне документи, які містять "кіт", але не містять "собака".

Простота реалізації булевої моделі є найголовнішою її превагою. Булева модель дозволяє виконувати прості та точні пошуки, де результати повертаються відповідно до вказаних у запиті умов. Однак, вона не враховує контекст і семантику слів, що може призводити до недостатньо точних результатів. До того ж пошукова система, що реалізує булеву модель пошуку вимагає від користувача знання мови спеціальних запитів заснованої на булевих операторах, тож використання булевої моделі може бути складним для користувачів, які не знайомі з логічними операторами. Також до недоліків можна віднести відсутність критеріїв оцінювання релевантності і, як наслідок, можливості ранжування знайдених результатів за цим критерієм.

У сучасних пошукових системах, булева модель часто комбінується з іншими моделями, такими як векторна чи семантична, для забезпечення більш точних та релевантних результатів пошуку.

Розширений булевий пошук

З метою подолання недоліків простої булевої моделі пошуку, зокрема для забезпечення можливості ранжувати множину знайдених результатів, які

видаватимуться на виході пошукової ситсеми, була здійснена спроба ускладнити та удосконалити булеву модель. В результаті цього було запропоновано розширену булеву модель пошуку [55].

Розширений булевий інформаційний пошук - це покращена версія традиційної булевої моделі, яка дозволяє використовувати додаткові функції та можливості для знаходження більш точних та релевантних результатів.

Основні особливості розширеного булевого пошуку включають:

1. Апроксимація дозволяє встановити відстань або близькість між термінами в пошуковому запиті. Наприклад, можна вказати, що два терміни повинні бути віддаленими один від одного не більше ніж на певну кількість слів.

2. Розширення запиту дозволяє автоматично розширити пошуковий запит шляхом додавання синонімів, споріднених слів або інших пов'язаних термінів. Це допомагає знайти більше варіантів, які можуть бути пов'язані зі шуканим контекстом.

3. Вагові коефіцієнти дозволяють надати різним термінам в пошуковому запиті різну вагу або важливість. Наприклад, можна вказати, що один термін є більш важливим за інший і має велику вагу при ранжуванні результатів.

4. Фразовий пошук дозволяє знайти документи, де терміни в пошуковому запиті зустрічаються в певній послідовності. Наприклад, можна шукати конкретну фразу або слова, які мають бути поруч одне з одним.

Ці додаткові можливості розширеного булевого пошуку дозволяють користувачам більш гнучко та точно формулювати свої пошукові запити, отримувати більш релевантні результати та зекономити час при пошуку інформації. Багато сучасних пошукових систем мають вбудовані функції розширеного булевого пошуку, що дозволяє використовувати ці можливості безпосередньо під час пошуку.

Пошук з використанням нечітких множин

Пошук з використанням нечітких множин – це підхід до інформаційного пошуку, який використовує нечіткі множини для представлення та обробки

нечіткої або невизначеної інформації [56, 57, 58, 59]. У цьому підході терміни, які використовуються в пошукових запитах та документах, можуть мати нечіткі або розмите значення.

Основні особливості пошуку з використанням нечітких множин включають:

1. Лінгвістичні змінні. Використання нечітких множин дозволяє враховувати лінгвістичні змінні, такі як "високий", "низький", "швидкий", "повільний" і т.д., які можуть бути важливими для користувача при формулюванні пошукових запитів.

2. Розмите співставлення. Замість точного співставлення термінів, використовується розмите співставлення, що дозволяє кожному терміну мати певну ступінь відповідності до запиту або документа.

3. Моделювання невизначеності. Пошук з використанням нечітких множин дозволяє моделювати невизначеність або нечіткість інформації, що може бути корисним при обробці суб'єктивних або неясних запитів.

4. Розширення запиту. Пошук з використанням нечітких множин може включати методи розширення запиту, які додають синоніми, споріднені слова або інші пов'язані терміни для поліпшення результатів пошуку.

Векторна модель

Векторна модель інформаційного пошуку є однією з основних моделей, яка використовується для знаходження та ранжування документів за їхньою відповідністю до пошукового запиту [60, 61]. Ця модель базується на математичних методах та використовує поняття векторів для представлення документів та запитів.

Векторна модель передбачає, що кожен документ та запит можна представити у вигляді вектора, де кожен термін або термінологічний терм представляється як вимір довжини вектора. Кожен вимір вектора відповідає терму, а значення виміру відображає кількість входжень терму в документ або запит.

Для порівняння векторів документів та запиту використовуються різні метрики схожості, такі як косинусна схожість. Ця метрика вимірює кут між

векторами та визначає ступінь схожості між ними. Чим ближче значення косинусної схожості до 1, тим більш схожі документ та запит.

Векторна модель має кілька переваг:

1. Гнучкість полягає в тому, що векторна модель дозволяє використовувати різні метрики схожості та методи ранжування для знаходження найбільш відповідних документів до запиту.

2. Контекстуальний пошук полягає в тому, що векторна модель враховує контекст термінів, оскільки використовує весь пошуковий запит, а не окремі слова.

3. Розширення запиту полягає в тому, що в екторна модель дозволяє використовувати методи розширення запиту, такі як розширення за синонімами або спорідненими словами, для поліпшення результатів пошуку.

Однак, векторна модель також має свої обмеження, такі як проблема з розумінням семантики та контексту термінів. Також, великі колекції документів можуть вимагати значних обчислювальних ресурсів для виконання пошуку.

У сучасних пошукових системах, векторна модель часто комбінується з іншими моделями, такими як булева чи семантична, для забезпечення більш точних та релевантних результатів пошуку.

Ймовірнісний пошук

Ймовірнісний інформаційний пошук (Probabilistic Information Retrieval) [62] – це підхід до інформаційного пошуку, який використовує ймовірнісні методи для знаходження та ранжування документів за їхньою відповідністю до пошукового запиту.

У ймовірнісному інформаційному пошуку використовується статистична модель, яка оцінює ймовірність того, що документ відповідає запиту. Ця модель базується на ідеї, що документи та запити можна розглядати як випадкові події, і ймовірність відповідності документу запиту визначається за допомогою статистичних методів.

Основні особливості ймовірнісного інформаційного пошуку включають:

1. Модель ймовірності. Використовується статистична модель, яка враховує ймовірність зустрічі термінів в документах та запитах. Ця модель дозволяє ранжувати документи за ймовірністю їхньої відповідності запиту.

2. Зворотний документний частотний рахунок (IDF). Ймовірнісний пошук враховує частоту зустрічі термінів в документах, що визначається за допомогою IDF. Це дозволяє визначити важливість термінів та відфільтрувати загальні або надмірно вживані слова.

3. Модель релевантності. Використовується модель релевантності для оцінки ймовірності, що документ відповідає запиту. Ця модель може враховувати різні фактори, такі як частота зустрічі термінів, довжина документа, кількість входжень запиту в документ і т.д.

4. Розширення запиту. Ймовірнісний пошук може використовувати методи розширення запиту, що додають синоніми, споріднені слова або інші пов'язані терміни, щоб поліпшити результати пошуку.

1.6.Комп’ютерно-лінгвістичний підхід

Особливістю інформаційного простору є наявність багатосторонніх та багатопрофільних джерел інформації. Сукупність потоків від цих джерел – величезна послідовність інформаційних повідомлень та документів на сторінках вебсайтів, соціальних медіа та соціальних мереж – утворює інформаційне середовище, що забезпечує як споживання та накопичення, так і розширене відтворення інформації. Значна частина інформаційних потоків в межах яких здійснюється інформаційний пошук є джерелами даних, що зберігаються у текстовому вигляді. Виявлення потрібної інформації, можливо навіть, невідомої та прихованої у контексті набору текстових документів, формування бази знань (БЗ) та, зокрема, набору ключових слів, що однозначно характеризують зміст документу, може здійснюватися за допомогою інтелектуального аналізу текстів (англ. Text mining). Інтелектуальний аналіз текстів або, іншими словами, глибинний аналіз текстів – це комп’ютеризований підхід до кількісного аналізу текстів, що охоплює методи машинного навчання, обробки природної мови,

статистичної класифікації, а також представляє собою набір технік для виявлення нетривіальних тенденцій та створення нової інформації (так званої інформації про інформацію) [63]. Text mining використовує різні методології аналізу, серед яких основне місце посідає обробка природної мови (Natural language processing – NLP), що застосовується для спеціального лінгвістичного аналізу й допомагає машині, так би мовити, «читати» текст. В основі обробки природної мови лежать теоретичні знання та методи обчислювальної лінгвістики (математичної, або ще іншими словами – комп'ютерної лінгвістики) як галузі комп'ютерних наук (Computer Science), що використовує математичні моделі для дослідження природних мов. Цьому міждисциплінарному напряму досліджень присвячені роботи як вітчизняних так і зарубіжних вчених. Серед вітчизняних науковців можна відзначити праці таких вчених, як Ланде Д. В., Широков В. А., Клименко Н. Ф., Палагін О. В., Шаронова Н. В., Висоцька В. А, Карпіловська Є. А., Комарова Л. І., Тарануха В. Ю. [64], Дарчук Н. П. [65], які займалися питанням автоматичної обробки природної мови та зробили значний внесок у розвиток української комп'ютерної лінгвістики [66]. Найбільший внесок серед зарубіжних вчених у галузі комп'ютерної лінгвістики зробили Хомський Н. [67], Reese R. M. [68], Jurafsky D. [69], Goldberg Y. [70].

Та в процесі використання, розуміння природної мови та її комп'ютеризованої обробки потрібно враховувати також й ряд проблем, які пов'язані, в першу чергу, з багатозначністю природної мови– полісемією. Адже природна мова містить різні форми слова (словоформи, що мають спільну основу), похідні від іншого слова, та мовні вирази, які використовуються для вираження різного змісту; тож їхнє значення в конкретній ситуації залежатиме від контексту [71]. Така мова ще називається переплетеною мовою (з англ. Inflected Language). Так, наприклад, виокремлюють наступні типи неоднозначностей [72], такі як:

– синтаксична або структурна (коли у процесі обробки природної мови не зрозуміло, про що саме йде мова у тексті, як, наприклад, про «слово» чи про «горобця» йдеться у прислів'ї «Слово – не горобець, вилетить – не впіймаєш»);

- смислова, значеннєва або лексична (коли в залежності від поставленого наголосу змінюється значення слова, як наприклад у слові «замок»);
- відмінкова (коли в залежності від відмінка слово «черговий» у реченнях «Український спортсмен виборов черговий трофей» та «Черговий приступив до роботи на вахті» виражає ознаку та особу відповідно);
- референційна (коли комп'ютер, у якого відсутнє, так би мовити, розуміння реальності не може встановити смислове відношення слова, наприклад займенника «неї» у фразі «Відкрий ліву полицю та візьми тарілку, я хочу в неї покласти фрукти» до «полиці» чи «тарілки»). Референційна неоднозначність, або іншими словами, некомпозиційність викликана відсутністю в природній мові правил, які б дозволяли визначити точне значення складного висловлювання не знаючи контекст, хоч і знаючи значення всіх інших складових слів у висловлюванні, адже деякі фрази можуть тлумачитись двояко.
- літерація (коли, наприклад, питання та відповідь в діалозі виражені нестандартно «Можеш відчинити вікно? – Мені холодно.». В той час, коли відповідь може бути дана у формі «так/ні»).

В основі вищезгаданих проблем лежать лінгвістичні складнощі, пов'язані зокрема зі синонімією, омонімією, стійким поєднанням слів та морфологічними варіаціями. Проблема синонімії пов'язана з тим, що одне поняття може мати декілька подібних за значенням слів. Граматичні омоніми – це слова різні за значенням, але однакові за написанням в окремих граматичних формах. Це можуть бути слова однієї або різних частин мови. Лексичні омоніми – це слова однієї частини мови, однакові за звучанням і написанням, але різні за лексичним значенням. Існують також стійкі словосполучення, що можуть мати зміст, який відрізняється від змісту кожного слова окремо. Також у багатьох природних мовах слова мають кілька морфологічних форм, що відрізняються за написанням. Та все ж незважаючи на проблеми, методи обробки природної мови достатньо добре (точність систем вирішення мовної неоднозначності складає 60-70% [72]) можуть

впоратися з покладеними на них задачами, тому є дуже корисними під час роботи з текстовими даними, які пов’язані з певною предметною галуззю.

Тож якщо говорити про побудову термінологічних онтологій, як один із видів формалізації предметних галузей на основі текстових інформаційних потоків [11, 15 ,16], якому присвячена ця дисертаційна робота, то важливо, щоб елементи цієї формальної схеми – терміни (слова та словосполучення), які використовуються в якості назв концептів, що супроводжують обрану предметну галузь, підпорядковувались принципу однозначності: слово, що використовується в якості назви, має бути назвою тільки одного об’єкта, якщо це одинична назва; а якщо це загальна назва, то це словосполучення має бути загальною назвою для всіх об’єктів одного класу. Наприклад, слова та словосполучення, що використовуються у тексті у лапках, зазвичай мають інше значення, ніж те, яке подано в словнику, тож і розглядати їх під час побудови онтології потрібно у іншому значенні (для прикладу, кінотеатр “Жовтень”, книга “Чорний лебідь” тощо). Також власні назви, що зустрічаються у тексті й збігаються із загальновживаними словами за написанням, можуть мати інший зміст (для прикладу, вулиця Миколи Шпака, проспект Перемоги тощо). Також важливо зберегти дійсну семантико-синтаксичну структуру речення під час його детального формалізованого представлення у вигляді, що стане зрозумілим комп’ютеру та придатним для подальшої автоматизованої обробки.

Загалом, задачі, що пов’язані з обробкою природномовних текстів можна розподілити за рівнями [73]:

1. Задачі, що виникають на рівні роботи з цілим корпусом текстових документів – пошук дублікатів, інформаційний пошук.
2. Задачі, що виникають на рівні документа – аналіз тематики, пошук протиріч, побудова анотації документа (реферату).
3. Задачі, що виникають на рівні окремих абзаців – виявлення відношень між реченнями та словами, визначення мови та емоційної тональності тексту (сентимент-аналіз).

4. Задачі на рівні речень – синтаксичний розбір, токенізація та видалення стоп-слів.
5. Задачі на рівні фраз та словосполучень – виокремлення окремих слів, виявлення іменованих сутностей.
6. Задачі на рівні слів – морфологічний аналіз, розмічування частин мови (англ. Part-of-Speech tagging, PoS tagging), стемінг, лематизація, векторизація.
7. Задачі, що виникають на знаковому рівні організації природномовного тексту – зміна регістру букв, видалення пунктуації та пробілів.

Також у зв'язку зі складністю природної мови, як зазначено у роботі [74], аналіз знакового рівня організації природномовного тексту у практичному плані обмежений проблемою виокремлення синтаксичних розділових знаків від слова, виділенням абревіатур, скорочень тощо. У зв'язку з неоднозначністю трактування деяких послідовностей символів (наприклад, крапка чи тире) може виникати завдання виявлення та виправлення спотворених слів. Тож автори зазначають про необхідність розроблення знакового рівня організації тексту як початкового етапу побудови моделі розуміння тексту. Така попередня обробка тексту природної мови в залежності від задачі та конкретної реалізації може складатися зокрема з переведення всіх букв в тексті в нижній або верхній регістри, видалення цифр (чисел) або заміна на текстовий еквівалент (зазвичай використовуються регулярні вирази); видалення знаків пунктуації (зазвичай реалізується як видалення з тексту символів зі заздалегідь заданого набору) та видалення символів пробілів (англ. whitespaces).

Для вирішення завдання обробки природної мови існують різні підходи, зокрема основними вважаються:

– лінгвістичний підхід, що складається з граматичного, морфологічного, синтаксичного, семантичного, реферaційного та структурного рівнів аналізу текстових даних;
– статистичний підхід, що базується на статистичному зважуванні найбільш уживаних слів тексту без урахування контексту;

– символічний підхід, що базується на логіці, правилах та словниках розроблених людиною.

У сучасних методах обробки текстів природньої мови використовується не тільки апарат лінгвістики для аналізу текстів, а й статистичні методи, математична логіка і теорія ймовірностей, кластерний аналіз, методи штучного інтелекту, а також технології управління даними.

1.7.Статистичний аналіз

Виокремлення ключових слів із великого об’єму тестової інформації – складна для людини задача. А враховуючи постійний приплив нової інформації, в умовах обмеженості людських ресурсів, ця задача стає зовсім непосильною. Тож для вирішення завдання обробки, зважування та виокремлення ключових термінів все частіше застосовуються автоматичні системи. Не дивлячись на різницю в методах, велика їх частина намагається зробити приблизно одне й те ж саме: використовуючи деяку евристику або числову міру (наприклад, відстань між словами або частоту використання слів), знайти групу слів, яка точно визначають тему чи описують інформацію, яка міститься у вхідному тексті.

Огляд існуючих методів статистичного зважування термінів показав, що для вирішення завдання виокремлення ключових термінів із текстових документів існує безліч методів, і постійно розробляються нові рішення. Лінгвостатистичні методи дозволяють аналізувати великі обсяги текстових даних, виявляти частотність вживання слів, фраз, конструкцій та інших мовних одиниць, а також встановлювати тенденції та зміни у вживанні мови.

Одним з основних інструментів лінгвостатистичного аналізу є розрахунок частотних характеристик, таких як частота вживання слів, типологічні характеристики (наприклад, типологічна частота), статистичні міри асоціації (наприклад, взаємна інформація) та інші. Ці характеристики дозволяють виявляти особливості функціонування мови, встановлювати зв'язки між словами та конструкціями, а також визначати стилістичні та жанрові особливості текстів.

В основі статистичного підходу лежить припущення, що зміст тексту відображається за допомогою найбільш уживаних у тексті слів. Тож ідея статистичного аналізу полягає у підрахунку кількості входжень або повторень певного слова у текстовому документі та обчисленні його вагового значення. Існують різні способи обчислення вагових значень термінів. Найпростіший з них – це коли вважається, що вагове значення дорівнює кількості зустрічань терміна t в документі d. На початку свого зародження статистичний підхід був зосереджений на аналізі частоти використання слова в тексті, але більша їх частина розглядала ключові слова у відношенні тільки одного документа та не враховувала дискримінаційну силу слова. Тільки в 70-х роках стала більш популярна ідея статистично аналізувати частоту вживання слів у документі по відношенню до великої кількості інших документів. Зокрема для вирішення цієї задачі й до тепер активно використовується статистичний ваговий показник *TF-IDF* (з англ. Term Frequency – частота слова, Inverse Document Frequency – обернена частота документа), завдяки якому оцінюється важливість слів у контексті документа, що є частиною колекції документів чи корпусу [75].

Термін *TF* введений Карен Спарк Джонс [76] для обчислення відношення числа входжень обраного слова до кількості слів у документі. У такий спосіб оцінюється важливість слова t_i в межах обраного документа:

$$\boldsymbol{TF} = \frac{\boldsymbol{n_i}}{\sum_{\boldsymbol{k}} \boldsymbol{n_k}}, \tag{1.1}$$

де n_i – число входжень слова t_i 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.

$$\boldsymbol{IDF} = \log \frac{|\boldsymbol{D}|}{|(\boldsymbol{d_i} \supset \boldsymbol{t_i})|}, \tag{1.2}$$

де $|D|$ – кількість документів колекції; $|(d_i \supset t_i)|$ – 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*:

$$TF\text{-}IDF = TF \circ IDF. \quad (1.3)$$

Вага (значимість) слова пропорційна кількості вживань цього слова у документі і обернено пропорційна частоті вживання слова у інших документах колекції. Більшу вагу *TF-IDF* отримають слова з високою частотою появи в межах документа та низькою частотою вживання в 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 [77].

Наприклад, більш ефективний метод, що заснований на статистичному аналізі, є латентно-семантичне індексування (англ. Latent Semantic Indexing (LSI)) [78, 79]. Латентно-семантичний аналіз використовується для отримання контекстно-залежних значень слів за допомогою статистичної обробки великих корпусів текстів. В основі LSI лежить матричний аналіз, який дозволяє знаходити семантичні зв'язки між словами та документами у текстовому корпусі.

LSI використовує метод сингулярного розкладу (Singular Value Decomposition, SVD) [80] для зменшення розмірності матриці термін-документ та створення нових векторів, які представляють семантичні концепти. Ці вектори можна використовувати для порівняння та кластеризації документів, а також для пошуку документів за семантичною подібністю.

Однією з переваг LSI є здатність до врахування семантичних зв'язків між словами, навіть якщо вони не зустрічаються в одному контексті. Таким чином, LSI може виявляти схожість між документами, які містять різні слова, але виражають подібні концепти.

Крім того, LSI може використовуватися для побудови рекомендаційних систем, де враховується семантична подібність між користувачами та предметами. Наприклад, на основі LSI можна рекомендувати користувачеві подібні фільми,

книги або товари на основі аналізу їх семантичного змісту. Таким чином, латентно-семантичне індексування є ефективним методом, що дозволяє виявляти семантичні зв'язки та зменшувати розмірність текстових даних для подальшого аналізу та використання.

Істотним недоліком методу є обчислювальна швидкість, де зі збільшенням обсягу вхідних даних (наприклад, SVD-перетворенні) спостерігається значне зниження швидкості обчислення.

Для вирішення завдання зважування та виокремлення ключових термінів також використовується дисперсійна оцінка важливості термінів [81]. Обчислення виконуються наступним чином: нехай деякий термін A позначається як A_k^n, де $k = 1,2, \dots, K$ – кількість зустрічань даного терміна у тесті, а n – позиція даного терміна у тексті. Для прикладу, A_4^{30} буде означати, що на 30-й позиції у тексті знаходиться термін A, який зустрівся четвертий раз. Величина $\Delta A_k = A_{k+1}^m - A_k^n = m - n$ – це інтервал між послідовними появами терміна, де на m-й позиції в тесті знаходиться термін A, що зустрічався k+1-й раз й на та n-й – k разів.

Дисперсійна оцінка важливості терміна розраховується наступним чином:

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

де $\langle \Delta A \rangle$ – середнє значення послідовності $\Delta A_1, \Delta A_2, \dots, \Delta A_K$, $\langle \Delta A^2 \rangle$ – середнє значення послідовності $\Delta A_1^2, \Delta A_2^2, \dots, \Delta A_K^2$, K – кількість зустрічань терміна A у тексті.

Основним недоліком представлених статистичних підходів є відсутність можливості врахувати зв'язки у тексті. А представлення тексту у вигляді множини слів є недостатнім для відображення його змісту, оскільки текст, насправді, є послідовним набором слів, що підпорядковуються певним правилам. Для подолання вказаних недоліків пропонується використовувати статистичні підходи у поєднанні з лінгвістичними підходами до аналізу тексту.

1.8.Лінгвістичний аналіз текстів

Відомо декілька етапів лінгвістичного підходу обробки текстових даних для автоматичного виокремлення ключових термінів, що включає:

1. Графемний (граматичний) аналіз текстових даних.
2. Морфологічний аналіз текстових даних.
3. Синтаксичний аналіз текстових даних.
4. Семантичний аналіз текстових даних.

Результат застосування кожного рівня використовується в якості вхідних даних для наступного рівня опрацювання й аналізу.

1.9.Синтаксичний аналіз

Синтактика – це розділ лінгвістики та семіотики, що вивчає відношення між знаками в рамках знакової системи. Синтаксичний аналіз (або синтаксичний розбір) [82] є одним з найскладніших етапів лінгвістичного методу обробки текстових даних. В процесі синтаксичного аналізу формуються синтаксичні групи або так звані синтаксичні структури – сукупності залежних слів або слів із групами, які показують, який зв’язок існує між словами. Синтаксичний аналіз здійснюється за допомогою синтаксичного аналізатора (парсера, англ. parser) із застосовуванням спеціальних фіксованих синтаксичних правил – граматики мови. В результаті послідовність слів у тексті подається у вигляді деревоподібної ієрархії, яка відображає синтаксичну структуру речення та синтаксичні залежності слів у реченні. Структура у вигляді дерева є зручним представленням, яке стає придатним для подальшої обробки комп’ютером.

Основна проблема синтаксичного підходу полягає у складній структурі речень у флективних мовах, таких як українська, російська, польська, на відміну від англійської, італійської, французької (аналітичні мови), де структура речень значно простіша. І як наслідок, для деяких природних мов зі складною морфологічною структурою дуже складно розробити чітку граматику, яка б її описувала. Тож зазвичай використовуються підходи, які використовують декілька

типів граматик. Та при збільшенні кількість слів у реченні й збільшенні кількості синтаксичних правил, експоненційно збільшується й складність алгоритму. І як наслідок процедура автоматичного синтаксичного аналізу стає трудомісткою для обчислень.

1.10. Семантичний підхід

Семантика – це розділ семіотики і лінгвістики, в рамках якого вивчаються знаки і знакові системи, як засоби вираження значення і змісту. Основним предметом семантики є інтерпретації знаків і їх поєднання. Семантика, як правило, розглядається в рамках міждисциплінарної області досліджень знаків і знакових систем семіотики спільно з двома іншими її розділами: синтактикою і прагматикою. Як вже було зазначено, перша з них вивчає відношення та взаємозв’язки знаків між собою (синтаксис), інша – відношення між знаками та суб'єктами, які з ними асоціюються та інтерпретуються, в той час як семантика розглядає знаки у їх відношенні до об'єктів (не мають знакової природи), які ними позначаються.

Тож додатковим кроком, який виконується після синтаксичного аналізу є семантична обробка [83], що полягає у визначенні для кожного слова чи словосполучення певного семантичного класу. На відміну від синтаксичного підходу, що вивчає синтаксичну структуру, семантика означає значення слів та словосполучень, а також речень загалом. Під семантичним аналізом розуміється інтерпретація речень у межах певного контексту. Цей крок дає змогу зрозуміти, чи відповідає слово або речення за змістом певному контексту. Як результат структура вхідного тексту перетворюється у внутрішню модель представлення даних певної автоматизованої системи – онтологію. Такими онтологічними представленнями можуть виступати «семантичні мережі», «семантичні графи» або «семантичні карти».

Семантичні мережі – це модель зображення знань, яка використовує графічне представлення для відображення семантичних відношень між різними сутностями

[84, 85, 86]. Такий підхід базуються на теорії графів і використовуються для аналізу семантики інформації.

Семантичні мережі можуть бути побудовані на основі відповідних текстів, що означає, що вони можуть використовувати тексти як джерело інформації для виявлення семантичних зв'язків між словами, фразами або реченнями [87]. Цей підхід дозволяє збагачувати модель знань і покращувати розуміння текстів за допомогою семантичного аналізу.

Використання семантичних мереж на основі відповідних текстів має декілька переваг. По-перше, вони дозволяють автоматично виявляти семантичні зв'язки між різними елементами тексту, що спрощує аналіз та обробку інформації. По-друге, вони можуть бути використані для автоматичної анотації текстів, що дозволяє швидше знаходити інформацію та зрозуміти її семантику. По-третє, вони можуть використовуватися у семантичному пошуку для покращення пошукових запитів, допомагаючи знаходити більш точні результати.

Однак, варто відмітити, що побудова семантичних мереж на основі відповідних текстів може бути складною задачею, оскільки потребує великої кількості даних і обчислювальних ресурсів. Також, цей підхід може бути обмежений в плані точності, оскільки інтерпретація семантики тексту може бути суб'єктивною.

У світі інформаційних технологій семантичні мережі отримані на основі відповідних наборів текстів активно використовуються у таких областях, як машинне навчання, обробка природної мови, побудова інтелектуальних агентів та інші. Вони допомагають покращити розуміння текстів, забезпечуючи більш точну та ефективну обробку інформації.

Робота семантичних алгоритмів полягає у порівнянні семантичної структури нових документів (їх семантичних моделей) із вже наявною збереженою у базі «еталонною» семантичною структурою.

Семантична структура документа включає в себе семантичні моделі, які представляють семантичні зв'язки між різними елементами документа, такими як слова, фрази або речення. Ці моделі можуть бути побудовані на основі алгоритмів

обробки природної мови та аналізу тексту, зокрема завдяки алгоритму побудови направлених зважених мереж термінів, запропонованого у цій дисертаційній роботі [16].

Порівнюючи семантичну структуру нових документів з «еталонною» семантичною структурою, семантичні алгоритми можуть визначати семантичну подібність або відмінність між ними. Це дозволяє виявляти схожі теми, концепти або ідеї у нових документах, а також виявляти відмінності в змісті.

Семантичні алгоритми можуть бути використані для різних завдань, таких як автоматична категоризація документів, пошук схожих документів, рекомендації контенту або виявлення плагіату. Вони допомагають автоматизувати процес аналізу та обробки великих обсягів текстової інформації, що дозволяє ефективно знаходити та використовувати потрібну інформацію.

Але оскільки комп'ютер, який використовує інтелектуальну обробку текстів, не може насправді «розуміти» контекст тексту, то представлення не можна назвати цілком семантичним. Проте врахування цього важливого обмеження дає нам змогу з певним наближенням говорити про успішність представленого підходу.

1.11. Семантичний пошук

Сучасні інформаційно-пошукові системи використовують обмежений набір методів та підходів для пошуку інформації. Як результат, такі системи у відповідь на користувацький запит видають зазвичай видають набір окремих текстових документів. У той же час велика частина інформації, що представлена в Інтернеті та локальних сховищах даних, якими, наприклад, є електронні бібліотеки та архіви, виявляється недоступною через неефективність роботи наявних пошукових систем. Це пов'язано, зокрема і з тим, що такі пошукові системи використовують примітивні пошукові методи та підходи, що базуються на ідеї виявлення та виокремлення ключових слів, без врахування контексту та семантики окремих слів, які входять до запиту. Для врахування контексту та семантики інформаційного запиту під час інформаційного пошуку найбільш сучасним та перспективним

підходом є використання семантичних мереж, побудованих на основі відповідних текстів.

Семантичний пошук, як вид автоматизованого повнотекстового інформаційного пошуку, є одним із найперспективніших напрямків розвитку інформаційно-пошукових систем. Ідея семантичного пошуку полягає в урахуванні семантичних зв’язків, які існують між словами та словосполученнями користувацького запиту та зв’язків у реченнях текстових документів тих інформаційних ресурсів, що були проіндексовані пошуковою системою. Семантичний інформаційний пошук - це підхід до пошуку інформації, який базується на розумінні семантики тексту, а не просто на співставленні ключових слів. Він використовує семантичні технології, такі як аналіз природної мови (Natural Language Processing, NLP) і семантичне моделювання, для розуміння змісту тексту та знаходження зв'язків між різними елементами. Цей метод дає змогу знайти ті документи, які навіть не містять слів із пошукового запиту, проте пов’язані з ними за змістом.

Перші спроби реалізації семантичного пошуку розпочалися не так давно, в кінці XX століття. Наприклад, у 2000 році P. Vakkari [88] запропонував спосіб пошуку схожих за семантикою документів на основі зіставлення їх лексичних векторів.

Для досягнення семантичного інформаційного пошуку використовуються різні технології і методи:

1. Аналіз природної мови (Natural Language Processing, NLP). NLP – це галузь штучного інтелекту, яка займається обробкою та розумінням природної мови людей. Використовуючи методи NLP, семантичний інформаційний пошук може розуміти семантику тексту, виявляти синоніми, антоніми, контекстуальні зв'язки та інші семантичні відношення між словами та фразами.

2. Векторне представлення слів (Word Embeddings). Це техніка, яка перетворює слова на числові вектори, які відображають семантичні відношення між словами. Застосування таких моделей, як Word2Vec або GloVe, дозволяє вимірювати семантичні відстані між словами і знаходити семантично близькі

слова. Векторне представлення слів дозволяє виміряти семантичну подібність між словами та знаходити семантичні зв'язки у тексті.

3. Моделі глибокого навчання (Deep Learning Models). Глибоке навчання використовується для побудови семантичних моделей, які можуть розуміти та виявляти семантичні зв'язки у тексті. Нейронні мережі, такі як рекурентні нейронні мережі (RNN) і трансформери, можуть бути використані для аналізу та моделювання семантики тексту.

4. Алгоритми машинного навчання. Машинне навчання використовується для навчання моделей, які можуть розуміти та аналізувати семантику тексту. Алгоритми, такі як класифікатори, регресія, кластеризація та ансамблі моделей, можуть бути використані для пошуку та аналізу семантичних зв'язків у тексті. Також застосування алгоритмів машинного навчання дозволяє системі навчитися розпізнавати семантичні відношення, такі як схожість, синонімічність, антонімічність тощо, і використовувати їх для пошуку інформації.

5. Онтології та бази знань. Також відомо про застосування онтологічного підходу для інформаційного пошуку. Онтології – це формалізовані моделі знань, які описують семантику домену або конкретної галузі. Онтологія також є методом подання та обробки знань і запитів, та використовуються для організації та представлення знань про об'єкти, поняття та їх взаємозв'язки. Також це модель концептуалізації предметної галузі, що призначена для опису семантики даних деякої предметної галузі та вирішення проблеми несумісності і суперечливості понять. В роботі [89] визначення онтології дано як: «явна специфікація концептуалізації, де в якості концептуалізації виступає з описом множини об'єктів і зв'язків між ними».

6. Семантичні бази даних та бази знань. Семантичні бази даних використовуються для зберігання інформації з урахуванням її семантики. Бази знань, які ґрунтуються на онтологіях, допомагають зрозуміти семантику тексту та знаходити зв'язки між різними елементами.

6. Семантичне моделювання. Семантичне моделювання використовується для побудови семантичних моделей, які представляють семантичні зв'язки між

елементами тексту. Ці моделі можуть бути побудовані на основі методів машинного навчання, статистики або експертних знань. Вони допомагають розуміти семантику тексту та знаходити зв'язки між різними елементами.

7. Семантичні алгоритми порівняння. Ці алгоритми використовуються для порівняння семантичної структури нових документів з вже існуючою "еталонною" семантичною структурою. Вони допомагають визначати семантичну подібність або відмінність між документами та допомагають знаходити релевантну інформацію.

В якості онтології, як моделі представлення текстової інформації, в процесі семантичного інформаційного пошуку може розглядатися семантична мережа, яка побудована на основі відповідних текстів. Це однин із найпоширеніших сучасних підходів до семантичного моделювання. Семантична мережа – це представлення тексту, яке включає в себе семантичні зв'язки між словами, фразами або реченнями. Вона дозволяє розуміти контекст і семантику тексту, а не просто виявляти наявність ключових слів. Це дозволяє забезпечити більш точні результати пошуку і забезпечити користувачеві більш релевантну інформацію. Тож одним з ключових аспектів семантичного інформаційного пошуку є побудова семантичних моделей – семантичних мереж, зокрема.

Ці мережі дозволяють автоматично виявляти семантичні зв'язки між різними елементами тексту і покращують розуміння текстів. Тож робота семантичних алгоритмів полягає у порівнянні семантичної структури нових документів (їх семантичних моделей – семантичних мереж) із вже наявною збереженою у базі "еталонною" семантичною структурою – "еталонною" семантичною мережею.

Одним з переваг семантичного інформаційного пошуку є його здатність розуміти семантику запиту користувача. Наприклад, якщо користувач шукає "ресторани в центрі міста", традиційна пошукова система може просто знайти всі документи, що містять слова "ресторан" і "центр міста". Однак, семантичний інформаційний пошук може розуміти, що користувач шукає ресторани, які знаходяться в центрі міста, і надати більш точні результати.

Семантичний інформаційний пошук має широкі застосування в різних галузях. Наприклад, він може бути використаний для покращення пошукових систем, рекомендаційного аналізу, аналізу соціальних мереж, обробки природної мови, розробки чат-ботів та інших додатків, які працюють з текстовою інформацією та завдань, пов'язаних з аналізом текстової інформації.

Однак, варто відзначити, що семантичний інформаційний пошук також має свої виклики. Побудова семантичних моделей вимагає великої кількості даних та обчислювальних ресурсів. Також, інтерпретація семантики тексту може бути суб'єктивною, і результати пошуку можуть залежати від використаних алгоритмів та методів.

Висновки до розділу 1

Здійснено огляд сучасного стану проблеми та наукових розробок, яким присвячена тема дисертації. Було розглянуто сучасні комп’ютерно-лінгвістичні підходи та методи автоматичного аналізу текстових інформаційних потоків з метою розпізнавання знань з предметної галузі з якою змістовно пов’язані текстові дані. Було встановлено, що існує декілька підходів, зокрема такі як статистичний та лінгвістичний. Здійснено огляд існуючих методів статистичного зважування термінів, серед них найбільш відомою й використовуваною в наш час оцінкою важливості термінів є TF-IDF. Встановлено, що статистичний показник важливості терміна TF-IDF показує наскільки добре даний термін визначає документ по відношенню до корпусу. Терміни з більш високим числовим значенням TF-IDF є важливими в межах певного документа й нечасто зустрічається в інших документах корпусу. Крім цього було акцентовано увагу й на проблемах, які можуть виникати під час використання методів статистичного зважування. Детально розглянуто основні рівні лінгвістичної обробки текстових даних. Розглянуто основні ідеї семантичного пошуку, як одного із найперспективніших видів автоматизованого повнотекстового інформаційного пошуку.

РОЗДІЛ 2. ТЕОРЕТИЧНІ ЗАСАДИ ФОРМУВАННЯ МЕРЕЖЕВОЇ МОДЕЛІ ПРЕДМЕТНОЇ ГАЛУЗІ

Цей розділ присвячено новим методам та підходам формування направлених зважених мереж термінів на основі тематичних інформаційних потоків, як семантичних моделей представлення текстових даних, що змістовно відносяться до певної предметної галузі. Запропоновано та представлено цілісну методику направлених зважених мереж із ключових термінів, як семантичних моделей предметних галузей на основі текстових корпусів. Зокрема представлено та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency), який в порівнянні зі звичайним статистичним показником TF-IDF дозволяє більш ефективно виокремити ключові слова та словосполучення тематичного текстового масиву. Також запропоновано новий метод виокремлення ключових термінів із текстового корпусу зі застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Також у цьому розділі здійснено огляд алгоритмів графів видимості (Visibility Graph algorithm – VG), що можуть використовуватись для формування мережевих моделей предметних галузей. Зокрема, досліджено можливість застосування алгоритму графа горизонтальної видимості для побудови ненаправленої мережі із ключових слів та словосполучень. Запропоновано та представлено нові правила визначення напрямків зв’язків між вузлами ненаправленої мережі, побудованої зі ключових термінів. Запропонований, розроблений та представлений новий метод визначення напрямків зв’язків у мережі термінів із застосуванням Part-of-speech tagging та новий підхід до визначення вагових значень цих зв’язків. Представлена апробація запропонованого статистичного показника важливості термінів, методу виокремлення ключових термінів, алгоритму побудови ненаправленої мережі із ключових слів та словосполучень, правил визначення напрямків зв’язків між вузлами ненаправленої мережі, підходу до визначення вагових значень цих зв’язків та цілісної методики та технологічної схеми формування направлених зважених мереж із ключових термінів на прикладі українського, англійського та китайського перекладів тексту священної книги Тори, П'ятикнижжя Мойсеєвого. Загалом було

опрацьовано всі п'ять книг – «Буття», «Вихід», «Левит», «Числа» та «Повторення закону».

2.1.Мережа ключових термінів

Оскільки на сьогоднішній день в інформаційних сховищах, розподілених у мережі, накопичуються терабайти неструктурованих або слабоструктурованих текстових даних, то для забезпечення пошуку розміщеної в мережі інформації необхідна розробка нових підходів та методів дослідження та структуризації цих даних [90], а також нових моделей їх представлення. При цьому, безумовно, повинні враховуватись переваги та недоліки вже існуючих моделей та алгоритмів, що застосовуються для інформаційного пошуку.

Сучасний розвиток інформаційних технологій дозволяє у деяких випадках знаходити необхідну інформацію в мережах. Проте залишаються невирішеними проблеми адаптації сучасних інформаційно-пошукових систем до потреб користувачів та підвищення пертинентності таких систем. Більшість із цих проблем – актуальні питання семантичної обробки величезних об’ємів динамічних текстових масивів [91] та побудови семантичних моделей. У загальному випадку семантичне представлення є графом, семантичною мережею, що відображає бінарні відношення між вузлами – словами та словосполученнями, що є смисловими одиницями тексту. Робота семантичних алгоритмів полягає у порівнянні семантичної структури нових документів (їх семантичних моделей) із вже наявною збереженою у базі «еталонною» семантичною структурою. Тож використання семантичних моделей під час інформаційного пошуку (семантичний пошук) дає змогу врахувати семантичні зв’язки, які існують між словами та словосполученнями користувацького запиту та зв’язки між ключовими термінами у реченнях текстових документів тих інформаційних ресурсів, що були проіндексовані пошуковою системою. Також опис семантики текстових даних предметної галузі в процесі інформаційного пошуку дозволяє вирішити проблеми,

що пов’язані з сумісністю або суперечливістю понять у текстах та семантичною відповідністю або суперечливістю текстів загалом.

Отже, під час комплексних досліджень певної проблемної предметної галузі, з якою тематично пов’язані текстові інформаційні потоки, важливим етапом є її концептуалізація та детальне формалізоване представлення знань (набору об’єктів та сутностей реального світу та зв’язків між ними) у вигляді, що придатне для подальшої автоматизованої обробки – побудова онтологічної моделі предметної галузі.

Виявилось, що багато задач, що виникають під час роботи з інформаційними потоками, мають багато чого спільного з лінгвістикою та математичними науками. Цей факт відкриває широкі можливості для застосування теорії лінгвістики та потужного математичного апарату [91, 13]. Лінгвістична теорія як розділ загального мовознавства, в свою чергу, дає змогу працювати з природномовними текстами, знаючи при цьому їх контекст, структуру, властивості та будову. З іншого боку, враховуючи проблеми розмірності та динаміки інформаційних ресурсів в глобальних мережах, для дослідження інформаційних потоків застосовується знання з області статистики, дискретної математики, зокрема теорії графів та складних мереж.

Тож в якості онтології в процесі семантичного пошуку зручною й ефективною мережевою моделлю представлення текстових даних, яка стандартизує інформацію у вигляді, що придатний для комп’ютерної обробки, може розглядатися семантична мережа. У назві сполучені терміни з двох наук: семантика у мовознавстві вивчає сенс одиниць мови, а мережа в математиці є різновидом графу – набору вершин, сполучених дугами (ребрами), що мають певне присвоєне їм вагове значення. Тож семантична мережа – це інформаційна модель предметної галузі, що має вигляд орієнтованого графу, вершини якого відповідають об'єктам предметної галузі, якими в свою чергу можуть бути поняття, події, властивості, процеси [93], а ребра задають відношення між ними. У семантичній мережі побудованій на основі текстових даних, що відносяться до певної предметної галузі, вузли відповідають окремим ключовим термінам (словам

та словосполученням) у тексті, що в свою чергу співвідносяться з об’єктами предметної галузі, а зв’язки – семантико-семантичним зв’язкам між цими ключовими термінами у тексті [94, 95]. Отже, тексти визначеної тематичної спрямованості можна представити у вигляді мережі із ключових слів та словосполучень, пов’язаних між собою формальним смисловим зв’язком, що дає змогу відобразити семантику предметної галузі, з якою тематично пов’язані тексти.

Процес побудови онтологій на основі величезного об’єму текстових корпусів зазвичай є складним та ресурсозатратним. Так, наприклад, під час побудови направлених зважених мереж із ключових термінів, як семантичних моделей представлення текстових даних, відкритою та до кінця не вирішеною проблемою є визначення та виокремлення базових об’єктів (ключових термінів – слів та словосполучень, що відповідають поняттям, подіям, властивостям та процесам предметної галузі). Також у зв'язку зі складністю природної мови, не менш складною й відкритою проблемою концептуалізації є встановлення відношень – семантико-синтаксичних зв’язків між вузлами мережі, що відповідають термінам, а також визначення та встановлення напрямків таких зв’язків та їх вагових значень. Не менш важливою є автоматизація вищезгаданих процесів.

Окремий крок такої формалізації – визначення базових об’єктів (в даному випадку – створення словникових номенклатур, тезаурусів та предметних словників з термінів, визначених на основі тематичних масивів текстових документів). Ефективний вибір окремих термінів й, тим більше, автоматизація такого відбору з текстового масиву – актуальна й невирішена задача [96, 97].

Також в результаті огляду існуючих мережевих моделей представлення текстових даних, було встановлено, що для побудови мереж термінів існує декілька підходів і 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 [98, 99], належать одному і тому ж абзацу [100], поєднані синтаксично [101, 102] або семантично [103, 104]. Семантичний аналіз полягає у виділенні семантичних зв’язків, формуванні семантичного представлення текстів. Одним з можливих варіантів семантичного

представлення є структура, що складається з «текстових фактів». Одним з можливих варіантів семантичного представлення є структура, що складається з «текстових фактів». В цій дисертаційній роботі застосовується семантичний аналіз в межах кожного речення, що називається локальним семантичним аналізом, а «текстовими фактами» виступають виокремлені ключові слова та словосполучення, що відповідають концептам предметної галузі, з якою змістовно пов’язані тексти.

2.2.Методика побудови направленої зваженої мережі термінів

У цій дисертаційній роботі пропонується методика побудови направленої зваженої мережі термінів [25, 31], як онтологічної моделі представлення текстових даних. Направлена зважена мережа термінів (англ. Directed Weighted Network of Terms – DWNT, або просто мережа термінів) – це семантична модель представлення текстових даних, де вузлами такої мережі є ключові терміни (слова та словосполучення) тексту, які використовуються як назви концептів певної предметної галузі, з якою змістовно пов’язані тексти, а направлені зв’язки – семантико-синтаксичні зв’язки та відношення між цими термінами.

Для побудови DWNT використовується процедура, яка складається з декількох кроків:

1. попередня ком’ютеризована обробка тексту;
2. виокремлення ключових термінів;
3. побудова DWNT:
   – встановлення ненаправлених зв’язків;
   – встановлення напрямків зв’язків;
   – визначення вагових значень зв’язків.

2.3.Попередня комп’ютеризована обробка тексту

Лінгвістична складова комп’ютеризованої обробки природномовних текстів є однією з центральних етапів інтелектуалізації інформаційних технологій. Врахування лінгвістичних особливостей елементів природної мови дає змогу

працювати з текстами не лише як з набором символів, а й, так би мовити, “сприймати й розуміти” структуру та контекст. Тож використання подальших комп’ютеризованих етапів обробки тексту є незамінним для подальшої їх формалізації.

Сегментація та токенізація тексту

На рівні графемного аналізу текстів відбувається обробка послідовності символів, які надходять на вхід, та виокремлення структури тексту: окремих розділів, параграфів, речень та слів. Для попередньої комп’ютеризованої обробки природньої мови у цій роботі застосовуються деякі найпоширеніші прийоми, що включають автоматичну сегментацію на окремі речення та подальшу токенізацію тексту – сегментацію вхідного тексту на елементарні одиниці (токени, лексеми) [53].

Токенізація (лексичний аналіз) [105] здійснюється для проведення попереднього лексичного аналізу за допомогою лексичного аналізатора, токенізатора або так званого сканера, що дозволяє розбити текст та перетворити послідовність символів (букв, цифр, пунктуації та пробілів) у послідовність елементарних одиниць – токенів (лексем). Зазвичай сканер – це одна із основних функцій парсера (синтаксичного аналізатора).

Під лексемою або, іншими словами, токеном прийнято розуміти певну форму слова (словоформу) як самостійну значеннєву одиницю, яку розглядають у сукупності всіх своїх можливих форм та значень. Токенізація є зазвичай початковим етапом обробки текстів, оскільки дає можливість працювати зі словом як з окремою сутністю, знаючи при цьому його контекст [53].

Тож лексичний аналіз можна розділити на два етапи: сканування, яке сегментує вхідний рядок у синтаксичні одиниці, які називаються лексемами, та класифікує їх у класи токенів; та обчислення, яке перетворює лексеми в оброблені значення.

Зазвичай лексичний аналізатор сприймає комбінацію токенів як знаки й не проводить ніяких додаткових операцій таких як, наприклад, перевірка, чи відповідає кожній відкритій дужці «(» закрита «)».

Part-of-Speech tagging

На рівні морфологічного аналізу також здійснюється визначення морфологічних ознак та характеристик слова та дослідження його можливих форм, і як наслідок слову присвоюються ряд атрибутів, зокрема частини мови, до якої слово належить, та ознакам, якими характеризується в середині цієї частини мови (рід, число, відмінок і т.д.). В цій дисертаційній роботі [17, 19, 20] в межах кожного речення після токенізації здійснюється розмічування частин мови (англ. Part-of-Speech tagging, PoS tagging) [106], що полягає у віднесенні слова в тексті до певної частини мови й присвоєні йому відповідного тега.

Розмічування частин мови, як один з перших етапів комп'ютерного аналізу тексту, зазвичай є наступним кроком обробки природної мови, який застосовується після токенізації. У корпусній лінгвістиці, розмічування частин мови (англ. Part-of-Speech tagging, PoS tagging) або іншими словами просто розмічування (англ. tagging) (рис. 2.1) – це процес віднесення слова в тексті або корпусі до певної частини мови, заснований як на його визначенні, так і на його контексті — тобто, на його зв'язку з суміжними і спорідненими словами у фразі, реченні, або абзаці.

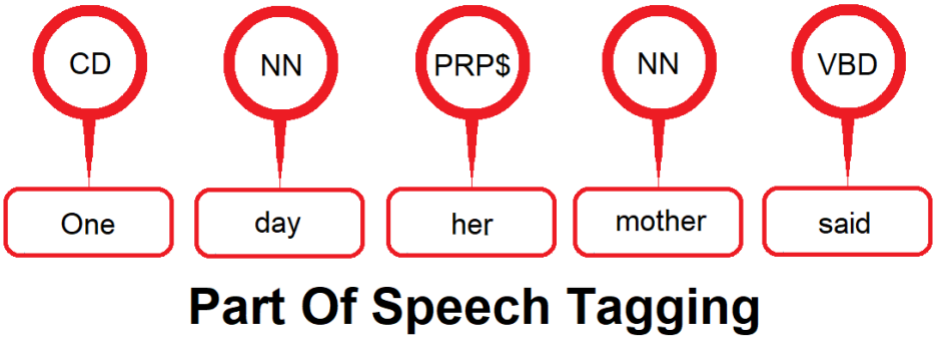

Рис. 2.1. Приклад розмічування частин мови

Також розмічування частин мови — одна із головних і базових складових практично будь-якого завдання NLP. Для розмічування частин мови використовується колекція наперед визначених тегів [107], які ставляться у

відповідність кожному слову в реченні. На рисунку 2.2 представлено набір тегів анотованого корпусу Treebank [106].

Tag	Part of speach	Tag	Part of speach
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	To
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WPS	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Рис. 2.2. Список тегів, що використані у Penn Treebank Project (без врахування знаків пунктуації) [107]

У лінгвістиці Treebank – це проаналізований та розмічений за частинами мови текстовий корпус, який коментує синтаксичну або семантичну структуру речень.

Оскільки частини мови також відомі як класи слів або лексичні категорії (які базуються на синтаксичному контексті фрази), то використання вищеназваного прийому класифікації слів за частинами мови дає змогу позначити кожне слово відповідно до його лексичної категорії. Тож PoS tagging може бути використаний для індексації слів, пошуку інформації і має також багато інших застосувань. Особливо PoS tagging може бути дуже корисним, якщо є слова або токени, які можуть мати декілька тегів. І найголовніше, розмічування спрощує контекст, який відноситься до певної предметної галузі.

Наприклад, одним із перших і найбільш широко використовуваних англійських розбірників на частини мови є метод розбору Е. Брілла [108], що

використовує алгоритми на основі правил. Окрім групи алгоритмів на основі правил існують і стохастичні алгоритми.

В цій дисертаційній роботі PoS tagging застосовується з мотою розділити слова або токени, які можуть мати декілька тегів, а отже – розділити слова, чиї нормальні словникові форми мають однаковий вигляд, але різне значення.

Лемматизація

Оскільки під час концептуалізації та побудови термінологічних онтологій предметних галузей на основі текстових документів визначеної тематики [16] важливо, щоб терміни (слова та словосполучення), які використовуються як назви концептів, що супроводжують обрану предметну галузь, підпорядковувалися принципу однозначності: слово, що використовується як назва, має бути назвою тільки одного об’єкта, якщо це одинична назва; а якщо це загальна назва, то це словосполучення має бути загальною назвою для всіх об’єктів одного класу. Тож в цій дисертаційній роботі додатково здійснюється лемматизація окремих розмічених лексем з метою отримати їх канонічні, словникові форми – леми. Лематизація – це приведення певної форми слова (словоформи), або іншими словами його лексеми до нормальної (базової, словникової) форми – леми. Лематизація здійснюється на початковому етапі обробки тексту після токенізації на елементарні одиниці. Однією із важливих особливостей лематизації є можливість визначення атрибутів тих слів, які відсутні у словнику. Так, наприклад, до деякого слова, якого немає у словнику, можна застосувати флективний (або флексійний) метод творення слів (творення похідних слів одного і того ж слова за допомогою флексії – шляхом зміни їхніх закінчень або звуків основи) й надати цьому незнайденому слову атрибути максимально підходящого.

Загалом існує два підходи морфологічного аналізу, що полягає у застосуванні побудованого словника слів, що містить їх характеристики, та використанні набору правил. Проте недоліком підходу на основі правил є недостатня точність результатів. Тож для вирішення задачі морфологічного аналізу та, зокрема, лематизації найбільш надійним та точним вважається використання підходу, що

базується на словнику слів. Зазвичай для пошуку інформації в текстах не потрібно проводити повний морфологічний аналіз, а достатньо лише перевірити, що деякі два слова є формами одного слова. Цей крок дозволяє додатково згрупувати різні форми одного й того слова, щоб їх можна було проаналізувати як єдиний елемент.

Існують різні програмні засоби та, зокрема, модулі NLTK (Natural Language Toolkit open source library) бібліотеки Python NLP (Natural Language Processing), що допомагають легко застосувати вищезгадані методи попередньої обробки до різних типів текстів.

Після розмічування тексту та отримання канонічних форм окремих лексем в межах кожного речення формуються ключові терміни [8].

Видалення стоп-слів

Щоб очистити текст від слів, що є джерелом шуму та не несуть ніякого додаткового смислового навантаження і є несуттєвими в контексті інформаційного пошуку, пропонується видалити стоп-слова. Наприклад, до стоп-слів належать артиклі прийменники, частки, вигуки, сполучники, прислівники, займенники, вступне слово, числа від 0 до 9 (однозначні), інші часто вживані службові, самостійні частини мови, символи, знаки пунктуації. Також до стоп-слів відносять часто використовувані в мережі Інтернет послідовності символів, як www, com, http та ін.

2.4.Статистичний показник важливості термінів GTF

Для вирішення завдання виокремлення ключових термінів із масиву текстових документів існує безліч методів, і постійно розробляються нові рішення. Одним із найбільш вживаних підходів є статистичний, в основі якого лежить припущення, що зміст тексту відображається за допомогою найбільш уживаних у тексті слів. Як наслідок, статистичний аналіз полягає у підрахунку кількості входжень або повторень певного слова у текстовому документі та обчисленні його вагового значення. Існують різні способи статистичного обчислення вагових

значень термінів. Найпростіший з них – це коли вважається, що вагове значення дорівнює кількості зустрічань терміна *t* в документі *d*.

Останнім часом все більш популярною є ідея статистично аналізувати частоту вживання слів у документі по відношенню до великої кількості інших документів. Зокрема для вирішення цієї задачі активно використовується статистичний ваговий показник *TF-IDF* (з англ. Term Frequency – частота слова, Inverse Document Frequency – обернена частота документа), завдяки якому оцінюється важливість слів у контексті документа, що є частиною колекції документів чи корпусу [75]. Вага або значимість слова пропорційна кількості вживань цього слова у документі і обернено пропорційна частоті вживання слова у інших документах колекції. Більшу вагу *TF-IDF* отримають слова з високою частотою появи в межах документа та в той же час низькою частотою вживання цих слів в інших документах колекції. Показник *TF-IDF* використовується в задачах аналізу текстів [1] та інформаційного пошуку. Також часто цей показник використовується як один з критеріїв релевантності документа до пошукового запиту [77].

Проте під час обробки тематичних інформаційних потоків потрібно враховувати той факт, що вони об’єднані певним спільним змістом, який може виражатися за допомогою однакових слів. Тож зрозуміло, що такі інформаційно-важливі слова будуть зустрічатися практично в кожному документі. А враховуючи той випадок, коли масив документів сформований за певним тематичним запитом (наприклад, запит із ключових слів, які повинні міститися у результатах пошуку), що й взагалі – інформаційно-важливе слово буде присутнє у кожному документі тематичного масиву. Як наслідок, використання вищезгаданого статистичного вагового показника *TF-IDF*, який враховує дискримінаційну силу слова, призведе до значного зниження ваги інформаційно-важливих слів, оскільки в межах кожного документа ці слова зустрічатимуться не дуже часто, а в межах всієї колекції документів навпаки – практично в кожному.

Звичайно, підхід, що полягає у використанні показника *TF-IDF*, дозволяє зменшити вагу таких часто вживаних, проте інформаційно-неважливих слів, як

артиклі, сполучники, прийменники. Але разом з тим використання цього вагового показника може призвести до втрати ключових слів, які зустрічаються дуже часто у тематичних інформаційних потоках. Проте якщо здійснити попередню обробку текстів, та видалити стоп-слова завчасно, то це відкриє можливості до застосування більш ефективних підходів та методів з точки зору обробки текстових документів однакової тематики.

Один з таких нових статистичних підходів був запропонований та досліджений у цій дисертаційній роботі та має назву глобальна частота терміна (англ. Global Term Frequency) або – *GTF* [1-3, 7-12]. Новий статистичний показник важливості термінів у тексті – *GTF* визначається відношенням загальної кількості появи терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу й показує, наскільки значимим є слово в глобальному контексті.

$$\boldsymbol{GTF} = \frac{\boldsymbol{n_i}}{\sum_k \boldsymbol{n_k}}, \tag{2.1}$$

де $\boldsymbol{n_i}$ – загальна кількість появи терма i у всіх документах корпусу; $\sum_k \boldsymbol{n_k}$– загальна кількість термів у документах корпусу.

Апробація запропонованого статистичного показника важливості термінів – глобальної частоти терміна GTF була проведена на прикладі загальновідомих біблійських текстів, які перекладені майже всіма мовами (зокрема, досліджувались тексти івритом, китайською, англійською, українською та російською мовами). Для обчислення статистичного показника важливості термінів й подальших досліджень був використаний англійський переклад тексту священної книги Тори, П'ятикнижжя Мойсеєвого, доступний за посиланням [109]. Зокрема було опрацьовано книгу «Числа», яка складається з 36 розділів.

Спочатку було здійснено попередню обробку кожного окремого розділу книги. Для попередньої обробки текстових даних застосовувалися деякі найпоширеніші прийоми, що включають автоматичну сегментацію на окремі речення та подальшу токенізацію тексту – сегментації вхідного тексту на елементарні одиниці (токени, лексеми) [53]. Як результат, текст кожного окремого

розділу книги був сегментований на речення та окремі слова – отримано послідовність лексем у тому порядку, в якому відповідні їм слова зустрічаються у тексті.

В межах кожного речення після токенізації також було здійснено розмічування частин мови (англ. Part-of-Speech tagging, PoS tagging) [106]. В результаті, окремі слова в тексті були віднесенні до певної частини мови й отримали відповідний тег. Далі серед всієї послідовності слів було виокремлено ті слова, які є іменниками (тег «NOUN») та власними назвами (тег «PROPN»). Для зручності виконання подальших кроків та, зокрема, формування шаблонів словосполучень, які будуть використані на етапі формування ключових термінів (слів, біграм та триграм), словам із тегом «PROPN» було присвоєно тег «NOUN».

Для нової послідовності, яка складалася лише з іменників та власних назв, додатково здійснювалася лемматизація окремих лексем з метою отримати їх канонічні, словникові форми – леми. Цей крок дозволив додатково згрупувати різні форми одного й того слова, щоб надалі їх можна було проаналізувати як єдиний елемент.

Для реалізації вищезгаданих етапів попередньої обробки текстових даних, сегментації, розмічування частин мови та лематизації, були застосовані програмні функції бібліотек мови програмування Python: pymorphy2 [110] – для текстів українською та російською мовою, NLTK (Natural Language Toolkit open source library) [111] – для англомовних тестів, Stanza [112] – для обробки текстів іншими доступними мовами (загалом Stanza підтримує 70 мов [113]).

Також щоб очистити текст від слів, що є джерелом шуму та не несуть ніякого додаткового смислового навантаження і є несуттєвими в контексті інформаційного пошуку, пропонується видалити ті іменники та власні назви, які є стоп-словам.

Після того, як для всіх 36-ти розділів були сформовані послідовності слів із тегом «NOUN», здійснювалося виокремлення 25-ти найбільш частотних слів в межах всієї книги. Тобто частота слова визначалася як загальна кількість зустрічань слова у всьому тексті – у кожному розділі.

Розділи книги розглядалися як окремі документи та були представлені у вигляді числових векторів, що відображають важливість використання 25-ти найбільш частотних слів у кожному розглянутому документі. Якщо слово не зустрічалося у певному розділі, то числове значення відповідної координати вектора дорівнює 0. Загалом було сформовано 36 векторів. Отримані для книги «Числа» числові вектори, значеннями яких є статистичний показник важливості термінів у тексті – глобальна частота терміна *GTF*, представлені у таблиці 2.1.

Таблиця 2.1

Векторне представлення книги «Числа» за допомогою статистичного показника важливості термінів *GTF*

	1	**2**	**3**	**4**	…	**33**	**34**	**35**	**36**
lord	0.052	0.052	0.052	0.052	**…**	0.052	0.052	0.052	0.052
child	0.034	0.034	0.034	0	**…**	0.034	0.034	0.034	0.034
son	0.031	0.031	0.031	0.031	**…**	0	0.031	0	0.031
mose	0.03	0.03	0.03	0.03	**…**	0.03	0.03	0.03	0.03
israel	0.03	0.03	0.03	0.03	**…**	0.03	0.03	0.03	0.03
offering	0	0	0	0.026	**…**	0	0	0	0
family	0.021	0.021	0.021	0.021	**…**	0.021	0	0	0.021
man	0.018	0.018	0.018	0	**…**	0.018	0.018	0.018	0.018
land	0.016	0	0.016	0	**…**	0.016	0.016	0.016	0.016
aaron	0.013	0.013	0.013	0.013	**…**	0.013	0	0	0
tribe	0.013	0.013	0.013	0.013	**…**	0.013	0.013	0.013	0.013
year	0.012	0	0	0.012	**…**	0.012	0	0	0
people	0	0	0	0	**…**	0.012	0	0	0
father	0.011	0.011	0.011	0.011	**…**	0	0.011	0	0.011
tent	0.01	0.01	0.01	0.01	**…**	0	0	0	0
congregation	0.01	0	0.01	0.01	**…**	0	0	0.01	0
priest	0	0	0.009	0.009	**…**	0.009	0.009	0.009	0
prince	0.009	0.009	0.009	0.009	**…**	0	0.009	0	0.009

lamb	0	0	0	0	**…**	0	0	0	0
ram	0	0	0	0	**…**	0	0	0	0
shekel	0	0	0.008	0	**…**	0	0	0	0
meal	0	0	0	0.008	**…**	0	0	0	0
house	0.008	0.008	0.008	0.008	**…**	0	0.008	0	0.008
levite	0.008	0.008	0.008	0.008	**…**	0	0	0.008	0
bullock	0	0	0	0	**…**	0	0	0	0

Також для порівняння було сформовано 36 векторів, числовими значеннями яких є статистичний показник важливості термінів *TF-IDF* (таблиця 2.2).

Таблиця 2.2

Векторне представлення книги «Числа» за допомогою статистичного показника важливості термінів *TF-IDF*

	1	**2**	**3**	**4**	…	**33**	**34**	**35**	**36**
lord	0	0	0	0	**…**	0	0	0	0
child	0.01	0.015	0.006	0	**…**	0.004	0.012	0.004	0.013
son	0.013	0.025	0.018	0.021	**…**	0	0.02	0	0.02
mose	0.002	0.002	0.005	0.004	**…**	0.002	0.002	0.001	0.004
israel	0.002	0.001	0.002	0	**…**	0.001	0.001	0.001	0.004
offering	0	0	0	0.003	**…**	0	0	0	0
family	0.061	0.008	0.081	0.072	**…**	0.005	0	0	0.059
man	0.004	0.002	0.001	0	**…**	0.001	0.002	0.002	0.001
land	0.002	0	0.002	0	**…**	0.025	0.025	0.032	0.004
aaron	0.005	0.003	0.018	0.021	**…**	0.006	0	0	0
tribe	0.041	0.038	0.002	0.002	**…**	0.003	0.077	0.004	0.087
year	0.048	0	0	0.038	**…**	0.007	0	0	0
people	0	0	0	0	**…**	0.003	0	0	0
father	0.032	0.009	0.009	0.012	**…**	0	0.006	0	0.038
tent	0.004	0.007	0.01	0.025	**…**	0	0	0	0

congregation	0.012	0	0.002	0.002	…	0	0	0.014	0
priest	0	0	0.013	0.006	…	0.003	0.004	0.014	0
prince	0.005	0.053	0.01	0.004	…	0	0.032	0	0.005
lamb	0	0	0	0	…	0	0	0	0
ram	0	0	0	0	…	0	0	0	0
shekel	0	0	0.033	0	…	0	0	0	0
meal	0	0	0	0.004	…	0	0	0	0
house	0.04	0.013	0.01	0.017	…	0	0.008	0	0.011
levite	0.026	0.016	0.05	0.008	…	0	0	0.031	0
bullock	0	0	0	0	…	0	0	0	0

Векторне представлення книги «Числа», що було виконане за допомогою статистичного показника важливості термінів *TF-IDF* у порівнянні з векторним представленням цієї книги виконаного за допомогою статистичного показника важливості термінів *GTF* значно відрізняється. Можна помітити, що наприклад, слово «lord», яке зустрічається у кожному окремому розділі книги «Числа» та, як наслідок, є найбільш частотним в межах всієї книги, безперечно, є також і ключовим в контексті цієї книги. Та зважаючи на те, що це слово є у кожному розділі, то числове значення статистичного показника *TF-IDF* дорівнюватиме 0. Те ж саме стосується й інших найбільш частотних слів, які є складовими майже кожного розділу – їх статистичний показник важливості у тексті *TF-IDF* дуже низький, або й взагалі дорівнює 0. Тож для документів, які мають спільну тематику, чи які пов’язані між собою – є складовими частинами певного загального документа – більш доцільним є використання *GTF* для виокремлення ключових слів.

Тож в результаті досліджень було показано, що запропонована оцінка важливості термінів *GTF* на відміну від звичайного статистичного показника *TF-IDF* дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно-важливий термін зустрічається майже у кожному документі корпусу.

2.5.Виокремлення ключових термінів

Кожне ключове слово потрібно правильно сформулювати. Мова ключових слів є умовною формалізованою мовою з використанням слів природної мови та арабських цифр. Існують певні правила визначення ключових слів та приведення їх до стандартної лексикографічної форми:

- ключове слово може бути виражене одним словом або словосполученням. Для цього використовують іменники, прикметники, прислівники;
- окремі прикметники не вважаються самостійними ключовими словами, проте вони можуть входити до словосполучення: коли характеризують іменник, наприклад (*біржові операції*); коли прикметник є невід'ємною частиною стійкого словосполучення, наприклад (*магнітні бурі*, *виробничі відносини*);
- поєднання іменника з іменником, якщо це широко поширені та стійкі, часто вживані словосполучення (наприклад, *джерела світла*, *засоби зв'язку*);
- словосполучення з одного іменника та кількох прикметників перетворюється на кілька двослівних ключових слів, в яких один і той же іменник супроводжується почергово одним прикметником (наприклад, світові фінансові кризи = *світові кризи*, *фінансові кризи*). У поєднаннях іменників з прикметниками інверсія не застосовується;
- у словосполученнях чисельника з іменником застосовується інверсія. Порядкові числівники інверсуються тоді, коли вони позначають черговість явища або події при їх послідовному повторенні, наприклад (*Олімпійські ігри*, *18-ті*).

Словосполучення виступають як ключові слова в наступних випадках:

а) якщо вони позначають назви різних законів, методів, теорій (*єдність та боротьба протилежностей*);

б) назви наукової дисципліни чи розділу науки (*економіка промисловості*)

в) назви різних типів організацій, партій, документів (*Організація Об'єднаних Націй*)

г) якщо один із компонентів словосполучення має надто широке термінологічне значення і тому доцільно використовувати його як самостійне слово через його неінформативність (*політичні партії*, *математичний апарат*);

д) якщо до складу словосполучень входить власне ім'я (*теорема Піфагора*, *закон Ома*).

Ключовими словами можуть бути:

– власні імена (наприклад, імена вчених, письменників, громадських діячів, такі як *Патон Б. Є.*, *Шевченко Т. Г.*);
– адміністративно-територіальні та географічні найменування (наприклад, *Житомирська область*, *Київ, Україна*);
– назви історичних подій, що наводяться у повній формі відповідно до наукової традиції (наприклад, *Галицька битва*, *Помаранчева революція*);
– хронологічні дані. Усі дати записуються арабськими цифрами, роки пишуться без «р», століття пишуться з «ст», наприклад (*XXI століття = 21 ст.*, *1991 рік = 1991*);
– слова, що використовуються для написання ключових слів, як правило, формулюються в називному відмінку і в множині, проте виняток становлять терміни, які не вживаються у множині (наприклад, *дихання*, *транспорт*, *ефективність*);
– в однині у формулюванні ключових слів вживаються назви окремих установ, організацій, індивідуальних предметів, власних імен і т. д. (наприклад, *Національна академія наук*, *Київський політехнічний інститут*);
– під час формулювання ключових слів велике значення має вживання повної чи короткої форми найменування предмета. Зазвичай надається перевага повній формі (наприклад, *Організація Об'єднаних Націй,* замість *ООН*), але якщо коротка форма витіснила повну назву предмета, то використовується коротка (наприклад, ЮНІСЕФ, ЮНЕСКО);
– однак найчастіше ж словосполучення вживаються в повній і скороченій формі (наприклад, інформаційно-пошукова система (ІПС), система підтримки прийняття рішень (СППР)).

В той час як існує багато методів виокремлення ключових термінів, що покладаються на частоту входження терміна (в документі, в корпусі, або їх комбінації), у метрик такого роду існують і деякі проблеми [114, 115], такі як залежність від корпусу і опора на припущення, що точне ключове слово має часто зустрічатися в документі, але рідко в інших документах корпусу. Ці методи так само ігнорують будь-які можливі відношення між словами в документі.

В цій дисертаційній роботі вважається, що ключовими термінами можуть бути окремі слова, які належать до такої частини мови, як іменник, а також словосполучення сформовані за наступними шаблонами:

- 2-грами: «ADJ~NOUN»;

- 3-грами: «NOUN~CCONJ~NOUN», «ADJ~ ADJ ~NOUN»;

де були використані слова, які відносяться до таких частини мови, як:

- іменник (тег NOUN) (зокрема загальним назвам (тег PROPN) для зручності було переприсвоєно тег NOUN);

- прикметник (тег ADJ);

- сполучник (тег CCONJ)).

Для вищезгаданої книги «Числа», що належить до священної книги Тори, П'ятикнижжя Мойсеєвого, для англійського перекладу тексту цієї книги було сформовано набір термінів (слів, а також словосполучень – біграм, триграм), у відповідності з вищезгаданими шаблонами та обчислено їх глобальну важливість за допомогою статистичного показника важливості термінів *GTF*. У таблицях 2.3, 2.4 та 2.5 наведено списоки 25-ти термінів (слів, біграм та триграм, відповідно) книги «Числа», що мають найвищі числові значення показника *GTF*.

Таблиця 2.3

Список 25-ти слів книги «Числа», що мають найвищі числові значення показника *GTF*

Слова	***GTF***
lord	0.0445
child	0.02881

son	0.0269
mose	0.02589
israel	0.02589
offering	0.02231
family	0.01805
man	0.01524
land	0.01334
aaron	0.0111
tribe	0.01087
year	0.01031
people	0.00986
father	0.00953
tent	0.00852
congregation	0.00852
priest	0.00807
prince	0.00785
lamb	0.00751
shekel	0.00695
meal	0.00695
house	0.0065
levite	0.0065
bullock	0.00617
meeting	0.00605

Таблиця 2.4

Список 25-ти біграм книги «Числа», що мають найвищі числові значення показника *GTF*

Біграми	***GTF***
fine~flour	0.00303

young~bullock	0.00247
sweet~savour	0.00202
continual~burnt	0.00191
silver~dish	0.00157
golden~pan	0.00157
silver~basin	0.00146
holy~thing	0.00112
common~man	0.00067
holy~convocation	0.00067
servile~work	0.00067
dead~body	0.00056
unleavened~bread	0.00045
red~sea	0.00045
open~land	0.00045
strong~drink	0.00034
cloud~abode	0.00034
evil~report	0.00034
midianitish~woman	0.00034
seventh~month	0.00034
high~priest	0.00034
holy~furniture	0.00022
holy~vessel	0.00022
unleavened~wafer	0.00022
cushite~woman	0.00022

Таблиця 2.5

Список 25-ти триграм книги «Числа», що мають найвищі числові значення показника *GTF*

Триграми	***GTF***

child~of~israel	0.01872
tent~of~meeting	0.00594
mose~and~aaron	0.00202
sacrifice~of~peace	0.00168
child~of~gad	0.00146
year~without~blemish	0.00146
child~of~reuben	0.00135
son~of~aaron	0.00135
land~of~canaan	0.00135
son~of~nun	0.00123
wilderness~of~Sinai	0.00112
hand~of~mose	0.00112
land~of~egypt	0.00101
plain~of~moab	0.00101
jordan~at~jericho	0.00101
city~of~refuge	0.00101
son~of~merari	0.0009
son~of~joseph	0.0009
son~of~ammihud	0.00078
son~of~gershon	0.00078
covering~of~sealskin	0.00078
son~of~levi	0.00067
son~of~jephunneh	0.00067
wilderness~of~zin	0.00067
water~of~sprinkling	0.00067

Природним та статистично зрозумілим є те, що глобальна частота *GTF* ключових словосполучень не перевищує глобальної частоти *GTF* його складових – слів, оскільки слова є складовими частинами словосполучень, і кількість зустрічань

у тексті ключових словосполучень не перевищує кількості зустрічань їх сладових (слів). Наприклад, слово «child» з глобальною частотою *GTF* рівною 0.02881 та «israel», із *GTF* 0.02589, входять до словосполучення «child~of~israel», яке має глобальну частоту *GTF* рівну 0.01872.

Для подальших кроків формування використовувались

2.6.Формування ненаправленої мережі термінів

Після виокремлення ключових термінів та їх статистичного зважування, в межах кожного окремого речення формується послідовність, де слова розташовуються у тому порядку, в якому вони зустрічаються у реченні тексту, а словосполучення з більшою кількістю слів розташовуються перед словосполученнями та словами, які є їхньою частиною. Для кожного речення тексту формується окрема послідовність термінів. Після цього здійснюється формування ненаправлених зв’язків між цими термінами.

Для побудови ненаправленої мережі для послідовності слів та словосполучень, у відповідність яким поставлені вагові значення в цій дисертаційній роботі використовується метод побудови графів видимості для ключових термінів (окремих уніграм/слів, біграм та триграм), зокрема – алгоритм побудови графа горизонтальної видимості (Horizontal Visibility Graph algorithm – HVG) [116, 117, 118].

Алгоритм графа горизонтальної видимості є розширенням стандартного алгоритму графа видимості (Visibility Graph algorithm – VG) [119, 120]. Цей алгоритм ставить у відповідність часовому ряду граф, сформований з його елементів (слів та словосполучень). Між вузлами, які відповідають елементам часового ряду, існує зв’язок, якщо вони знаходяться у “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) знаходяться в “прямій видимості”, а отже, є двома зв'язаними вузлами відповідного графа, якщо значення (t_c, y_c), що знаходиться між ними, задовольняє умову:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}. \quad (2.2)$$

Аналіз показав, що структура часового ряду зберігається у топології графу: періодичний ряд трансформується у регулярний граф, випадковий ряд – у випадковий граф, фрактальний ряд – у безмасштабний граф [119].

Зокрема, у цій дисертаційній роботі розглядається розширення стандартного алгоритму графа видимості [5, 6] – алгоритм графа горизонтальної видимості (Horizontal Visibility Graph algorithm – HVG), який використовується як один із етапів побудови мережевої структури на основі текстів, в яких окремим словам або словосполученням деяким чином поставлені у відповідність числові вагові значення. Відповідно до алгоритму графу горизонтальної видимості вважається, що два вузли t_i та t_j, які відповідають елементам часового ряду x_i і x_j, знаходяться у горизонтальній видимості тоді й тільки тоді, коли

$$x_k < min(x_i; x_j), \quad (2.3)$$

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

Побудова ненаправленої мережі термінів здійснюється з використанням вищеописаного алгоритму графа горизонтальної видимості. Спочатку 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.

У цій дисертаційній роботі в якості числових оцінок було використано запропонований статистичний показник важливості термінів у тексті – *GTF* (Global Term Frequency) – глобальна частота терміна, що визначається відношенням загальної кількості появи терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу й показує, наскільки значимим є слово в глобальному контексті. Було показано, що запропонована оцінка важливості термінів на відміну від звичайного статистичного показника *TF-IDF* дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно-важливий термін зустрічається майже у кожному документі корпусу.

Далі будується граф горизонтальної видимості. Таким чином, 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.

2.7.Встановлення напрямків зв’язків

У зв'язку зі складністю природної мови, відкритою проблемою концептуалізації та формалізації текстових даних, окрім встановлення семантико-синтаксичних зв’язків між вузлами мережі, що відповідають термінам, є також визначення напрямків таких зв’язків.

У цій дисертаційній роботі представлені нові методи та підходи до визначення напрямків зв’язків між вузлами ненаправленої мережі, побудованої зі слів та словосполучень тематичного текстового масиву [7, 8].

Напрямки зв’язків між вузлами, що відповідають термінам у тексті, визначаються відповідно до наступних запропонованих правил. 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 причинно-наслідковим зв’язкам між вузлами. Вважається, що $\forall_{i,j}$: $(t_i, t_j) \in T$ зв’язок існує у напрямку від вузла t_i до t_j якщо:

1. числове значення, що відповідає: а) степеню [121, 122] б) показнику, що обчислюється за алгоритмом HITS [123] в) показнику, що обчислюється за алгоритмом PageRank [124] вузла t_i більше за числове значення цього показника у вузла t_j;

2. у реченні термін, якому відповідає вузол t_i зустрічається раніше терміна, якому відповідає вузол t_j;

3. термін, якому відповідає t_i коротший за термін, якому відповідає t_j.

Для побудови направленої мережі зі слів та словосполучень (уніграм, біграм та триграм) за третім правилом використовується один із методів створення термінологічних онтологій – алгоритм формування мереж природніх ієрархій

термінів. Як зазначено у роботі [125] алгоритм створення мереж природніх ієрархій термінів передбачає побудову компактифікованого графу горизонтальної видимості, та встановлення направлених зв'язків між ключовими термінами, де напрямок зв'язку визначається за принципом входження слова у двочленне або тричленне словосполучення, та входження двочленного у тричленне.

Запропонований підхід для визначення напрямків зв'язків у ненаправлених мережах термінів був апробований на прикладі англомовного тексту, а саме – відомої казки "The story of Little Red Riding Hood" [126].

Відповідно до вищеописаного методу було здійснено обробку обраного текстового документу й виокремлено ключові терміни: уніграми, біграми та триграми (Таблиця 2.1).

Таблиця 2.1.

Топ-26 ключових уніграм для тексту "The story of Little Red Riding Hood" та їх степінь, HITS та PageRank.

Уніграми	GTF	Степінь	HITS	PageRank
grandmoth	0.065	49	0.444	0.0545
red	0.059	32	0.3099	0.036
hood	0.053	22	0.256	0.0252
ride	0.053	2	0.051	0.0029
wolf	0.031	30	0.301	0.0327
wood	0.025	17	0.204	0.0169
bed	0.019	18	0.191	0.0186
open	0.016	13	0.19	0.0148
time	0.016	14	0.199	0.0162
beauti	0.016	15	0.155	0.0154
big	0.012	9	0.088	0.0119
flower	0.012	8	0.126	0.0093
door	0.012	10	0.136	0.0125
cap	0.012	12	0.162	0.0141

cake	0.012	9	0.101	0.0116
wine	0.012	7	0.099	0.0094
cut	0.009	10	0.095	0.0127
mother	0.009	7	0.118	0.0092
strang	0.009	13	0.116	0.0167
jump	0.009	9	0.101	0.0116
weak	0.009	6	0.083	0.0083
ate	0.009	7	0.134	0.0081
live	0.009	6	0.097	0.0073
pull	0.009	7	0.12	0.0083
sick	0.009	4	0.05	0.0059
huntsman	0.009	11	0.113	0.0148

Таблиця 2.2.

Топ-22 ключові біграми для тексту “The story of Little Red Riding Hood” і їх степінь, HITS та PageRank.

Біграми	GTF	Степінь	HITS	PageRank
ride_hood	0.053	26	0.465	0.0277
red_ride	0.053	28	0.494	0.0303
grandmoth_big	0.009	11	0.177	0.0095
hood_wolf	0.006	5	0.179	0.0054
tasti_bite	0.006	8	0.192	0.0085
bed_pull	0.053	3	0.021	0.0038
hood_grandmoth	0.053	4	0.107	0.0047
leav_path	0.009	7	0.162	0.0084
snore_loudli	0.006	8	0.055	0.0068
grandmoth_live	0.006	6	0.164	0.0071
wolf_bodi	0.006	7	0.160	0.0080
pull_curtain	0.006	5	0.106	0.0056

grandmoth_bed	0.006	5	0.04	0.0060
straight_grandmoth	0.006	7	0.133	0.0076
sick_weak	0.006	5	0.068	0.0057
beauti_wood	0.006	6	0.133	0.0070
cake_wine	0.006	8	0.172	0.0084
beauti_flower	0.006	6	0.105	0.0074
wood_wolf	0.006	7	0.193	0.0076
press_latch	0.006	8	0.047	0.0064
grandmoth_sick	0.006	5	0.096	0.0058
door_open	0.006	8	0.109	0.0085

Таблиця 2.3.

Топ-26 ключових триграм для тексту “The story of Little Red Riding Hood” і їх степінь, HITS та PageRank.

Триграми	GTF	Степінь	HITS	PageRank
red_ride_hood	0.1429	36	0.67	0.1042
grandmoth_what_big	0.0252	6	0.145	0.0114
press_the_latch	0.0168	6	0.059	0.0111
leav_the_path	0.0168	4	0.153	0.0126
sick_and_weak	0.0168	6	0.182	0.0179
bed_and_pull	0.0168	7	0.162	0.0188
cake_and_wine	0.0168	5	0.160	0.0133
hear_how_beauti	0.0084	2	0.111	0.0076
look_so_strang	0.0084	2	0.03	0.0071
hood_and_ate	0.0084	2	0.111	0.0079
listen_littl_red	0.0084	2	0.129	0.007
bite_he_climb	0.0084	2	0.001	0.0101
obey_her_mother	0.0084	2	0.021	0.0084
lay_the_wolf	0.0084	2	0	0.0105

mind_your_manner	0.0084	2	0.054	0.0068
open_hi_belli	0.0084	2	0.003	0.0096
cake_and_drank	0.0084	2	0.018	0.0090
snore_veri_loudli	0.0084	2	0	0.0105
strang_oh_grandmoth	0.0084	2	0.028	0.0059
bird_are_sing	0.0084	2	0.021	0.0084
bed_fell_asleep	0.0084	2	0	0.0103
ride_hood_enter	0.0084	2	0.114	0.0074
larg_heavi_stone	0.0084	2	0.004	0.0093
woman_wa_snore	0.0084	2	0.111	0.0076
loudli_a_huntsman	0.0084	2	0	0.0105
red_ride_hood	0.0084	2	0	0.1042

Побудувавши направлену мережу за першим правилом було отримано наступні результати для різних числових показників вузлів мережі (а) для степеня вузла – рис. 2.1; б) для HITS – рис. 2.2; в) для PageRank – рис. 2.3). Візуалізація побудованих мереж термінів здійснювалась за допомогою засобів програмного забезпечення для моделювання та візуалізації графів – Gephi [127].

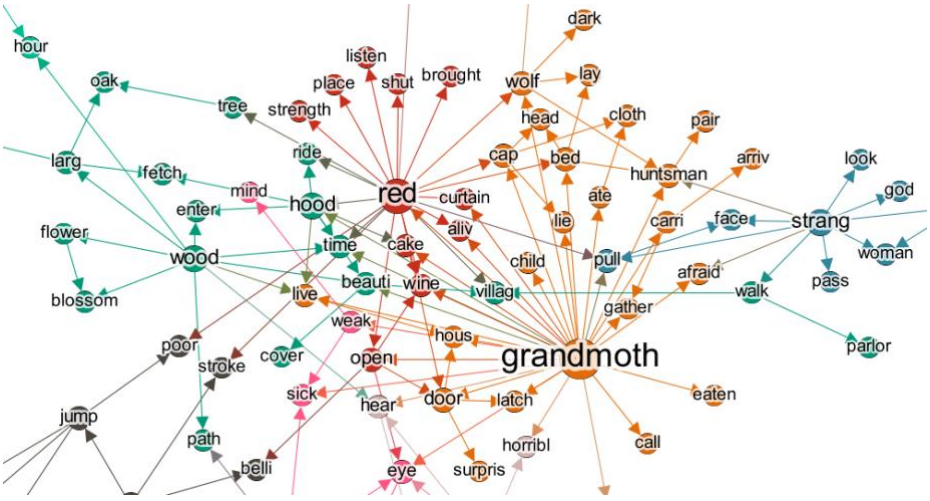

Рис. 1.1. Фрагмент направленої мережі побудованої за першим правилом для а) степеня вузла.

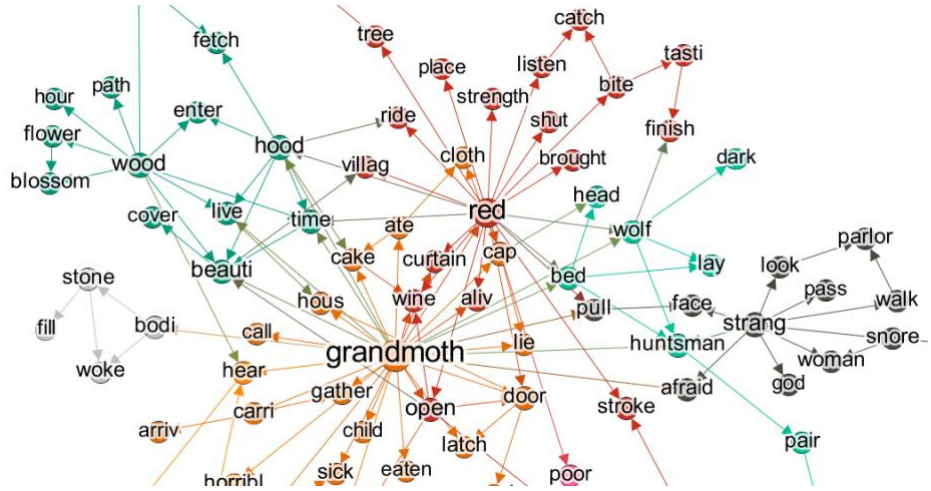

Рис. 2.2. Фрагмент направленої мережі побудованої за першим правилом для б) HITS.

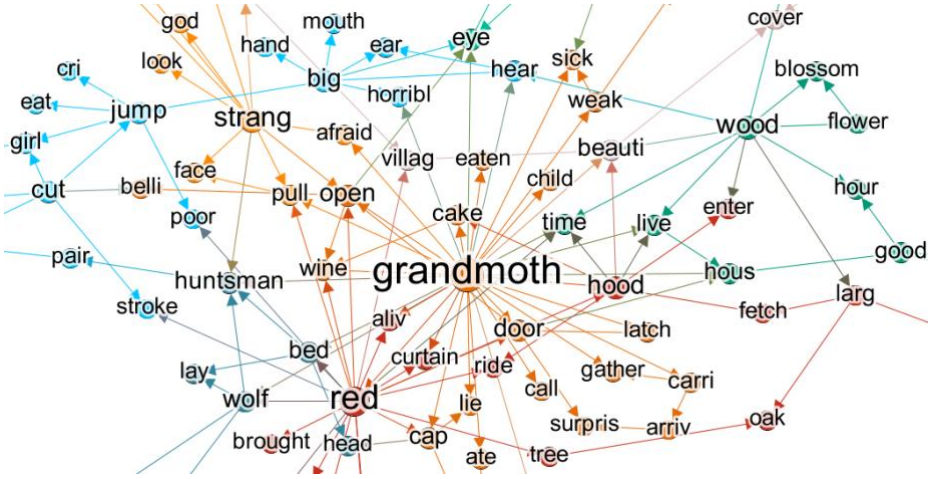

Рис. 2.3. Фрагмент направленої мережі побудованої за першим правилом для в) PageRank.

На рис. 2.4 представлена направлена мережа термінів, що побудована за другим правилом.

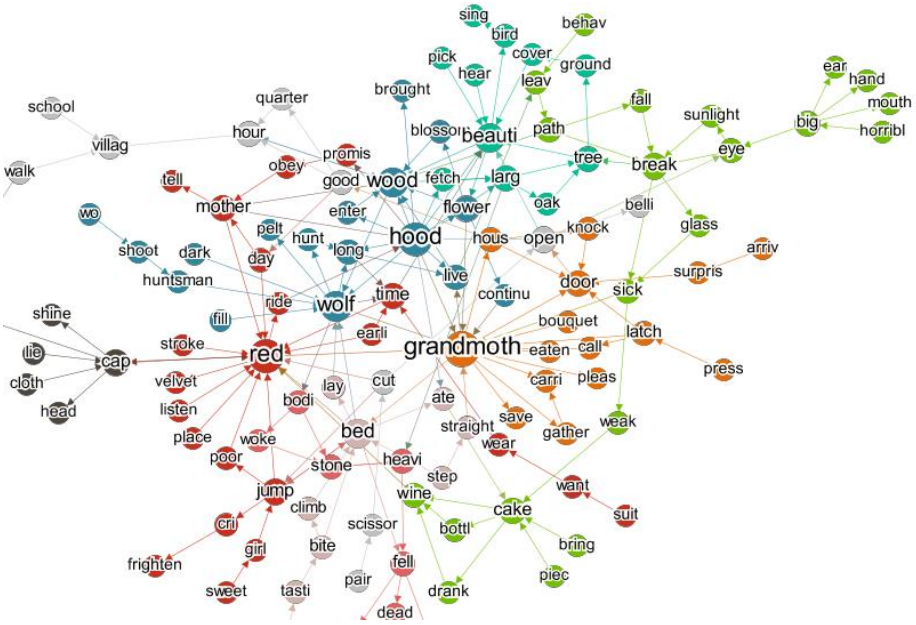

Рис. 2.4. Фрагмент направленої мережі побудованої за другим правилом.

На рис. 2.5 представлена мережа природніх ієрархій термінів, що побудована за третім правилом.

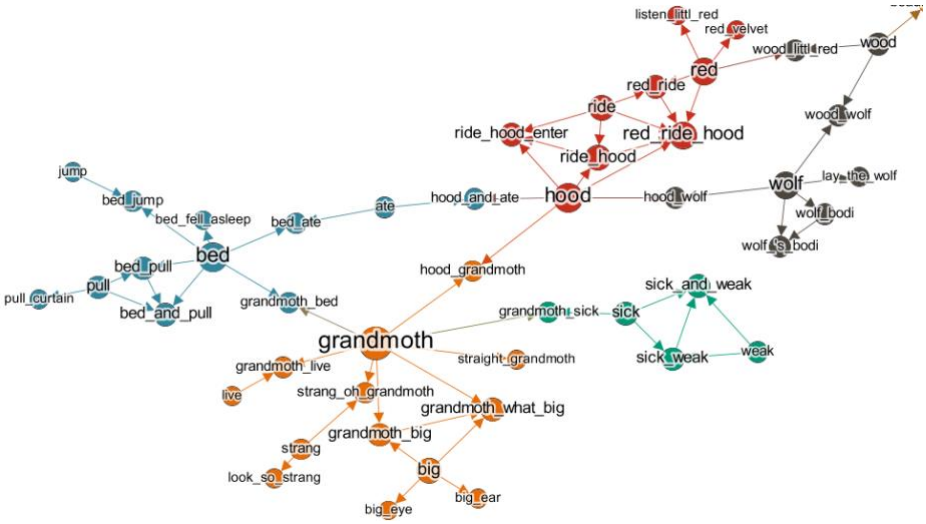

Рис. 2.5. Фрагмент мережі природніх ієрархій термінів.

Провівши аналіз отриманих результатів, було встановлено, що направлена мережа побудована за другим правилом більш точно відображає напрямок зв’язків, які існують між термінами у розглянутому тексті, ніж мережа, побудована за першим правилом. Мережа природніх ієрархій термінів має свої особливості та переваги, тому її важко порівняти із мережами, що побудовані за першими двома

правилами. Зважаючи на природність зв’язків, які встановлюються у такій мережі, можна говорити про їх синтаксичну адекватність.

Якщо розглянути, наприклад, напрямки зв’язків, що визначені для ключових термінів, то видно, що за першим правилом зв’язок між “wolf”-“grandmother”-“red” виглядає наступним чином (див рис. 2.1, 2.2 та 2.3): для степеня вузла, HITS та PageRank – “grandmother” впливає на “wolf” та “red”, а “red” на “wolf”, що не відповідає реальним напрямкам зв’язків у тексті з точки зору змістовного аналізу; в той час, як за другим правилом – “wolf” впливає на “grandmother”, а “grandmother” на “red”, що відповідає змісту тексту, який розглядався.

Тож правило визначення напрямків зв’язків, коли у реченні термін, якому відповідає вузол t_i зустрічається раніше терміна, якому відповідає вузол t_j $T(t_i, t_j) \in T$, є більш змістовним серед перших двох запропонованих, оскільки зв’язки, визначені за цим правилом більш точно відповідають змісту тексту на думку експертів.

2.8.Визначення вагових значень зв’язків

Для направленої мережі термінів, побудованої відповідно до попередніх кроків, в цьому розділі запропоновано та розроблено новий метод визначення вагових значень зв’язків, які встановлені між вузлами направленої мережі [9, 15, 16]. Загальний принцип на рівні графа описується наступним: 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 та відповідних рядків – зваженої компактифікації графа горизонтальної видимості. 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 зустрічань терміна t_i перед терміном t_j у тексті (кількістю зустрічань елемента часового ряду t_i перед елементом t_j).

Висновки до розділу 2

У цьому розділі запропоновано та представлено цілісну методику формування направлених зважених мереж із ключових термінів, як семантичних моделей предметних галузей на основі текстових корпусів. Зокрема запропоновано та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency) – глобальна частота терміна, що визначається відношенням загальної кількості появи терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу й показує, наскільки значимим є слово в глобальному контексті. Було показано, що запропонована оцінка важливості термінів на відміну від звичайного статистичного показника TF-IDF дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно-важливий термін зустрічається майже у кожному документі корпусу. Також запропоновано новий метод виокремлення ключових термінів із текстового корпусу зі застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Також у цьому розділі здійснено огляд алгоритмів графів видимості (Visibility Graph algorithm – VG), що можуть використовуватись для формування мережевих моделей предметних галузей, і запропоновано новий метод визначення напрямків зв’язків та їх вагових значень у мережі термінів.

РОЗДІЛ 3. ДОСЛІДЖЕННЯ ТА АНАЛІЗ МЕРЕЖ ТЕРМІНІВ

У цьому розділі здійснено огляд існуючих мережевих характеристик, що застосовуються для аналізу мереж та для дослідження особливостей її вузлів. Запропоновано алгоритм побудови динамічної мережі термінів та за його допомоги досліджено динаміку вагових значень вузлів у мережі термінів.

Також викладено методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж.

3.1.Алгоритми центральності

Алгоритми центральності використовуються, щоб оцінити та розуміти роль окремих вузлів у графі та їхнього впливу на мережу, яка визначається цим графом. Ці алгоритми корисні тим, що визначають найважливіші вузли та допомагають зрозуміти групову динаміку, наприклад довіру, доступність, швидкість поширення певних речей, а також визначити сполучні ланки (мости) між групами. Хоча багато з цих алгоритмів було винайдено для аналізу соціальних мереж, проте з тих пір вони знайшли застосування і в різних галузях і сферах діяльності.

3.2.HITS

Алгоритм ранжування HITS (Hyperlink Induced Topic Search) був запропонований та розроблений в 1998 році Дж. Клейнбергом (J. M. Kleinberg) [123] та забезпечує вибір із мережі, вузлами якої є документи, кращих «авторів» (першоджерел, на які посилаються інші документи) та «посередників» (документів, які посилаються на ці першоджерела). Задавши початкові значення важливості документа як «автора» $a_i^{(0)}$та «посередника» $h_i^{(0)}$, ітераційно обчислюються значення:

$$\boldsymbol{a_i^{(k)} = \sum_{j:e_{ij}\in E} h_j^{(k-1)}}, \quad \boldsymbol{h_i^{(k)} = \sum_{j:e_{ij}\in E} a_j^{(k-1)}}, \quad \boldsymbol{k = 1, 2, 3, \ldots} \tag{3.1}$$

3.3.PageRank

PageRank є найвідомішим з алгоритмів центральності [123]. Він вимірює транзитивний (або направлений) вплив вузлів. Усі інші алгоритми центральності, які згадувались вище, оцінюють прямий вплив вузла, тоді як PageRank враховує також вплив сусідів вузла та їхніх сусідів. Наприклад, наявність кількох дуже впливових друзів може зробити вас більш впливовим, ніж наявність багатьох менш впливових друзів. PageRank обчислюється або шляхом ітеративного розподілу рейтингу одного вузла серед його сусідів, або шляхом випадкового обходу графа, що відповідає мережі, та підрахунку частоти, з якою кожен вузол зустрічається під час цього випадкового обходу (блукання).

PageRank (Пейдж-ранк) – один з алгоритмів оцінки важливості та ранжирування вебсторінок за гіперпосиланнями, був створений в Стенфордському університеті Пейджем і Бріном в 1996 році й використаний в Google. PageRank названий на честь співзасновника Google Ларрі Пейджа, який створив його для ранжування веб-сайтів у результатах пошуку Google. Основне припущення полягає в тому, що сторінка з більшою кількістю вхідних і в той же час більш впливових вхідних посилань є більш ймовірним джерелом довіри. PageRank оцінює кількість і якість вхідних зв’язків із вузлом, щоб визначити, наскільки важливим є цей вузол. Передбачається, що вузли з більшим впливом на мережу мають більше вхідних зв’язків від інших впливових вузлів.

Інтуїтивно вплив полягає в тому, що зв’язки з більш важливими вузлами сприяють більшому впливу відповідного вузла, ніж еквівалентні зв’язки з менш важливими вузлами. Оцінка впливу зазвичай включає оцінку вузлів, часто зі зваженими зв’язками, а потім оновлення оцінок протягом багатьох ітерацій. Іноді оцінюються всі вузли, а іноді в якості репрезентативного розподілу використовується випадковий вибір.

Варто також зазначити, що показники центральності оцінюють важливість вузла порівняно з іншими вузлами. Тобто центральність – це рейтинг потенційного впливу вузлів, а не показник фактичного впливу. Наприклад, можна визначити

двох людей з найвищим центральним місцем у мережі, але, можливо, існують політичні чи культурні норми, які фактично і передають вплив на інших. Кількісна оцінка фактичного впливу є активною областю досліджень для розробки додаткових показників впливу.

В оригінальному документі Google PageRank визначається наступним чином:

$$\boldsymbol{PR(u) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)} \quad (3.2)$$

де:

- вважається, що сторінка u містить цитати зі сторінок T_1 до T_n (T_1 до T_n – сторінки, що посилаються (цитують) сторінку u);
- d – коефіцієнт демпфування, який встановлюється в інтервалі [0;1]. Зазвичай він встановлюється на 0,85. Ви можете вважати це ймовірністю того, що користувач продовжить ітераційні кроки. Цей коефіцієнт допомагає мінімізувати зниження рейтингу;
- $1 - d$ – це ймовірність того, що вузол буде досягнуто безпосередньо без переходу за будь-якими посиланнями (зв’язками);
- функція $C(T)$ визначається як вихідний степінь вузла T (дорівнює кількості посилань, що виходять зі сторінки T).

Для складної мережі, що задається матрицею суміжності $\hat{H}$, обчислюється $\hat{G}$:

$$\boldsymbol{\hat{G} = \alpha\hat{H} + [\alpha\bar{a} + (1 - \alpha)\bar{e}]\frac{1}{n}\bar{e}^T,} \quad (3.3)$$

де $\bar{a}$ – вектор-стовпець ($a_i = 0$ якщо з i-го вузла не виходить жодний зв'язок та $a_i = 1$– в протилежному випадку); n – кількість вузлів в мережі; α – коефіцієнт загасання (зазвичай $\alpha = 0.85$). Ліві власні значення $\hat{G}$ і є PageRank мережі.

3.4.Динамічні мережі термінів

В цьому розділі дисертаційної роботи також досліджується динаміка вагових значень вузлів у мережі термінів [34].

Нехай є деякий інформаційний потік d_i, $i = \overline{1,N}$, де N – кількість текстових документів у потоці. Алгоритм побудови динамічної мережі термінів складається з наступних кроків:

Крок 1. Нехай перший документ d_1 у потоці d_i є первинним для певної сукупності текстів T_i, що формується ($T_1 = d_1$). Для T_1 формується направлена зважена мережа термінів G_1, методика якої представлена у розділі 2. Отримана мережа термінів G_1 є первинною мережею, для якої фіксується список ключових термінів та їх вагових значень – *GTF*.

Крок 2. Сукупність текстів T_i, отримана на попередньому кроці, розширюється завдяки додаванню наступного документа з потоку d_i: $T_{i+1} = T_i + d_i$. В результаті буде отримано нову сукупність текстів T_{i+1} для якої формується нова направлена зважена мережа термінів G_{i+1} з урахуванням нових вагових значень виокремлених ключових термінів. Для отриманої мережі фіксується новий список ключових термінів та їх оновлених вагових значень – *GTF*.

Крок 3. Крок 2 повторюється, допоки $i \neq N$.

Далі для кожного ключового терміна із зафіксованого на останньому кроці списку ключових термінів будується графік динаміки вагового значення *GTF* окремого терміна в залежності від розширення сукупності текстів T_i на кожному кроці $i = \overline{1,N}$.

Алгоритм побудови динамічної мережі термінів був також апробований на прикладі загальновідомих біблійських текстів, які перекладені майже всіма мовами (зокрема, досліджувались тексти івритом, китайською, англійською, українською та російською мовами). Для обчислення динаміки вагових значень ключових термінів й побудови графіка динаміки був використаний англійський переклад тексту священної книги Тори, П'ятикнижжя Мойсеєвого, доступний за посиланням [109]. Зокрема було опрацьовано книгу «Числа», яка складається з 36 розділів. Кожен розділ книги «Числа» П'ятикнижжя Мойсеєвого роглядався як окремий документ інформаційного потоку d_i, $i = \overline{1,36}$ – кількість розділів у книзі «Числа».

В якості первинного документа d_1 було взято, відповідно, перший розділ книги. Тож на першому кроці було отримано сукупність текстів T_1, яка відповідає першому розділу, й для неї було сформовано направлену зважену мережу термінів G_1. У відповідності до первинної мережі G_1 було зафіксовано список ключових термінів та їх числових значень вагового показника *GTF*. Отримана для першого розділу книги «Числа» мережа термінів представлена на рисунку 3.1, а список 25-ти виокремлених ключових термінів та їх вагових значень *GTF* – у таблиці 3.1.

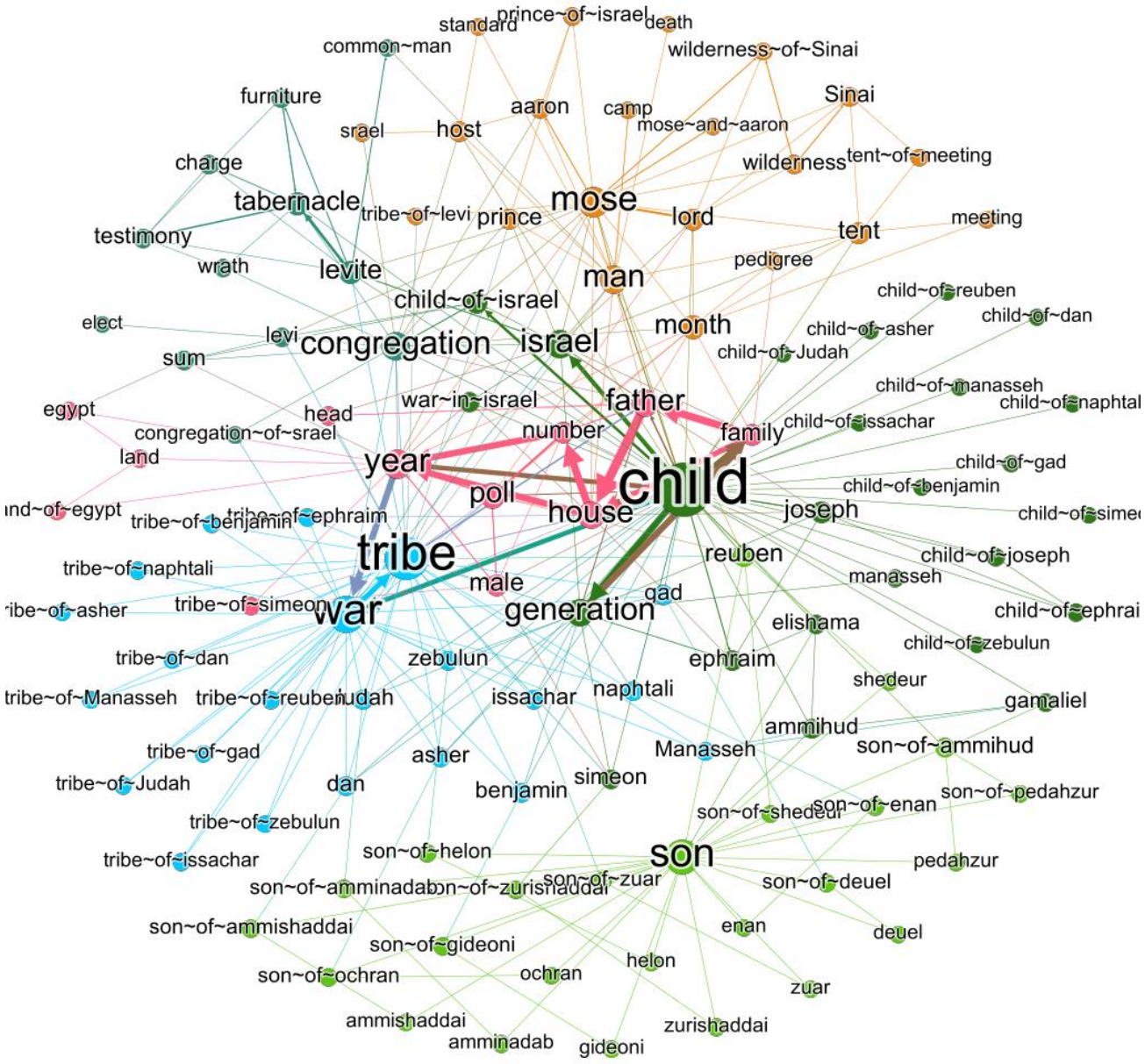

Рис. 3.1. Мережа термінів для першого розділу книги «Числа»

Таблиця 3.1

Список 25-ти термінів першого розділу книги «Числа», що мають найвищі числові значення показника *GTF*

Терміни	***GTF***
child	0.05764
father	0.05476
house	0.04899
tribe	0.04611
year	0.04611
family	0.04035
number	0.04035
war	0.03746
son	0.03458
generation	0.03458
israel	0.02882
man	0.02017
mose	0.01729
child~of~israel	0.01729
levite	0.01729
tabernacle	0.01729
congregation	0.01441
lord	0.01153
poll	0.01153
aaron	0.00865
male	0.00865
reuben	0.00865
gad	0.00865
Judah	0.00865
ephraim	0.00865

Відповідно до алгоритму побудови динамічної мережі термінів, для інформаційного потоку, сформованого з окремих розділів книги «Числа» П'ятикнижжя Мойсеєвого, було отримано 36 мереж термінів та відповідних їм числових векторів, що відображають важливість використання 25-ти термінів у кожній новій сформованій сукупності текстів T_i на кожному кроці $i = \overline{1, 36}$ (якщо на певному кроці у динамічно формованій сукупності текстів T_i термін ще не зустрічався, то числове значення відповідної координати вектора дорівнює 0). Список з 25-ти термінів був сформований на основі кінцевої сукупності текстів T_{36}, яка, в результаті застосування алгоритму побудови динамічної мережі термінів та розширення первинної сукупності текстів T_1, практично є об'єнанням всіх 36-ти розділів книги «Числа». Тобто сукупність текстів T_{36} є цілісною книгою «Числа» П'ятикнижжя Мойсеєвого, список ключових термінів (список слів, біграм та триграм, виокремлених з урахуванням числових значеннь статистичного показника *GTF*) якої представлений у таблицях 2.3, 2.4 та 2.5, відповідно. Загалом було сформовано 36 векторів. Отримані для книги «Числа» числові вектори, значеннями яких є статистичний показник важливості термінів у тексті – глобальна частота терміна *GTF*, представлені у таблиці 3.2.

Таблиця 3.2

Векторне представлення за допомогою статистичного показника важливості термінів *GTF* сукупностей текстів, сформованих із розділів книги «Числа»

	1	**2**	**3**	**4**	…	**33**	**34**	**35**	**36**
lord	0.0115	0.0127	0.0246	0.0235	**…**	0.046	0.0453	0.0444	0.0445
child	0.0576	0.0653	0.0535	0.0379	**…**	0.0274	0.0283	0.0281	0.0288
son	0.0346	0.0436	0.045	0.0477	**…**	0.0266	0.0272	0.0265	0.0269

mose	0.0173	0.0163	0.0246	0.0265	…	0.0268	0.0264	0.0259	0.0259
israel	0.0288	0.0254	0.0289	0.0212	…	0.0257	0.0254	0.0252	0.0259
offering	0	0	0	0.0008	…	0.024	0.0234	0.0227	0.0223
child~of~israel	0.0173	0.0181	0.0236	0.0167	…	0.018	0.0178	0.0179	0.0187
family	0.0403	0.0272	0.0385	0.0409	…	0.0187	0.0182	0.0177	0.018
man	0.0202	0.0163	0.0128	0.0091	…	0.0157	0.0155	0.0154	0.0152
land	0.0029	0.0018	0.0021	0.0015	…	0.0118	0.0124	0.0135	0.0133
aaron	0.0086	0.0073	0.0171	0.0227	…	0.0119	0.0116	0.0113	0.0111
tribe	0.0461	0.0436	0.0268	0.0197	…	0.0076	0.0095	0.0094	0.0109
year	0.0461	0.029	0.0171	0.0227	…	0.0111	0.0108	0.0105	0.0103
people	0	0	0	0	…	0.0106	0.0103	0.01	0.0099
father	0.0548	0.0399	0.03	0.0273	…	0.0088	0.0088	0.0086	0.0095
tent	0.0058	0.0073	0.0107	0.0189	…	0.0092	0.0089	0.0087	0.0085
congregation	0.0144	0.0091	0.0064	0.0053	…	0.0087	0.0085	0.0087	0.0085

priest	0	0	0.0064	0.0068	…	0.0081	0.008	0.0082	0.0081
prince	0.0058	0.0254	0.0203	0.0159	…	0.0074	0.0081	0.0079	0.0078
lamb	0	0	0	0	…	0.0081	0.0079	0.0076	0.0075
shekel	0	0	0.0054	0.0038	…	0.0075	0.0073	0.0071	0.0069
meal	0	0	0	0.0008	…	0.0075	0.0073	0.0071	0.0069
house	0.049	0.0363	0.0268	0.025	…	0.0065	0.0066	0.0064	0.0065
levite	0.0173	0.0145	0.0225	0.0174	…	0.0064	0.0062	0.0066	0.0065
bullock	0	0	0	0	…	0.0066	0.0065	0.0063	0.0062

У відповідності з результатами, що представлені у таблиці 3.2, для кожної окремої послідовності числових значень *GTF*, що відповідає певному терміну, було побудовано графік динаміки значень *GTF*. Наприклад, для ключового терміна «lord», після застосування алгоритму побудови динамічної мережі термінів, графік динаміки *GTF* цього терміна представлений на рисунку 3.2.

Рис. 3.2. Графік динаміки *GTF* терміна «lord» для різної сукупності текстів T_i, сформованої з розділів книги «Числа»

Використовуючи результати, представлені у таблиці 3.2 було отримано зведений графік динаміки *GTF* 10-ти найвагоміших ключових термінів – рисунок 3.2.

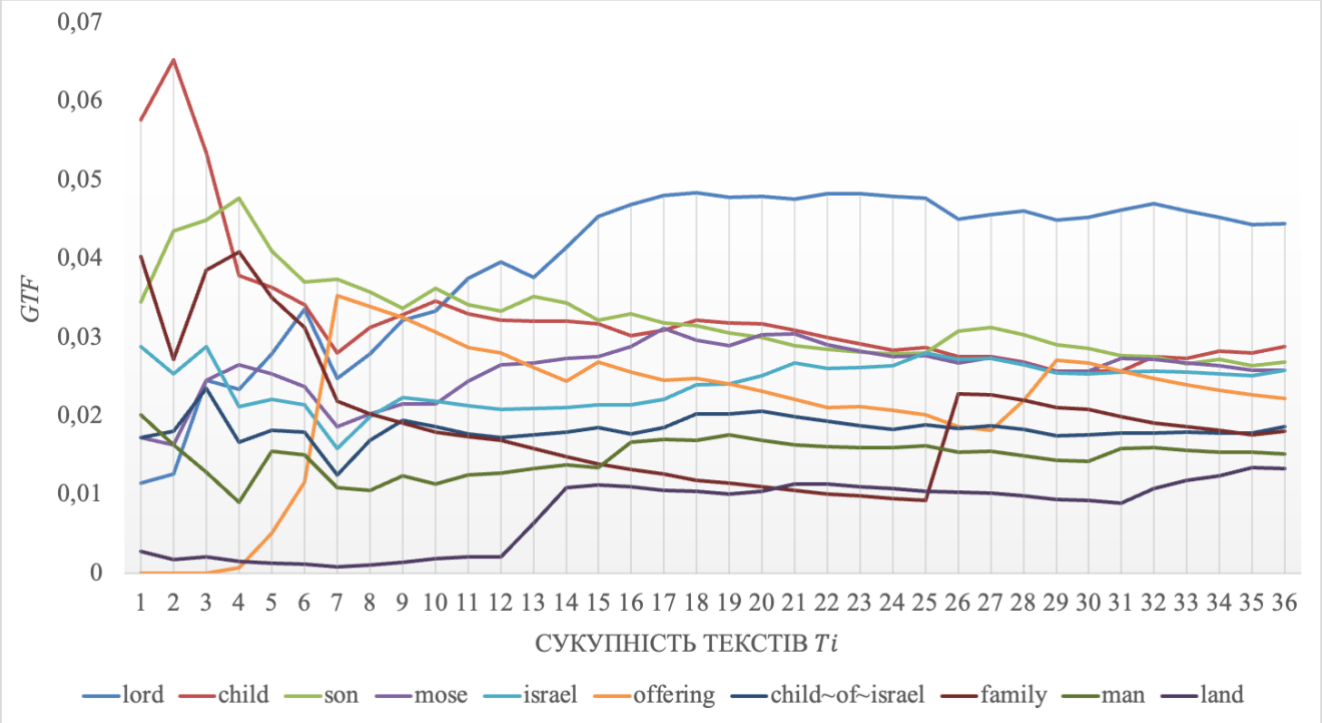

Рис. 3.2. Зведений графік динаміки *GTF* 10-ти найвагоміших ключових термінів для різної сукупності текстів T_i, сформованої з розділів книги «Числа» в результаті застосування алгоритму побудови динамічної мережі термінів

На рисунку 3.2 можна помітити, що деякі ключові терміни з'являлися, починаючи з деякого етапу i розширення сукупності текстів T_i. До певного етапу i числові значення показника важливості *GTF* цих термінів у результаті застосування алгоритму побудови динамічної мережі термінів дорівнювали 0. Наприклад, термін «offering» (оранджева лінія динаміки на графіку, що представлений на рисунку 3.2) вперше з'явився у сукупності текстів T_4 під час четвертого етапу ($i = 4$) розширення вищезгаданої сукупності – тобто під час розширення сукупності шляхом додавання розділу 4 книги «Числа» П'ятикнижжя Мойсеєвого. Також, поглянувши на динаміку терміна «offering» на графіку, представленому на рисунку 3.2, можна помітити, що цей термін активно використовувався в тексті впродовж 5-го, 6-го та 7-го розділів, оскільки спостерігається стрімке підвищення показника *GTF* цього терміна після додавання 5-го, 6-го та, особливо, 7-го розділів до розширюваної сукупності текстів T_i. Після цього, починаючи з 8-го розділу й до 27-го включно, спостерігалося поступове зниження числового значення *GTF* терміна «offering» з незначним підвищенням у 15-му та 18-му розділах. У 28-му та 29-му розділі знову помітне різке підвищення *GTF* цього терміна. Це свідчить про активну появу терміна «offering» у тексті цих розділів книги «Числа». Опісля, починаючи з 30-го та закінчуюючи 36-им розділом книги «Числа» П'ятикнижжя Мойсеєвого, термін «offering» не використовувався, а отже, значення *GTF* цього терміна знижувалось після розширення сукупності текстів за рахунок цих останніх розділів.

У такий спосіб, використовуючи алгоритм побудови динамічної мережі термінів, можна досліджувати динаміку окремих ключових термінів в результаті підвищення чи зниження їх глобальної частоти зустрічання у тексті шляхом додавання текстових документів, які насичені чи збагачені окремим визначеним терміном, у інформаційний потік. Такі термінологічні збагачення можуть бути

штучними й викликані «інформаційними вкидами», пропагандою чи спамом. Також вони можуть бути результатом навмисних, цілеспрямованих інформаційних атак – інформаційних операцій. Тож їх виявлення може бути здійснене шляхом аналізу динаміки ключових термінів, отриманої в результаті застосування алгоритму побудови динамічної мережі термінів.

Окрім динаміки глобальної частоти появи терміна, також досліджувалась динаміка інших показників вузлів, зокрема мережевих характеристик вузлів таких як *HITS* та *PageRank*, отриманих у результаті застосування алгоритму побудови динамічної мережі термінів. У якості апробації було використано той же англійський переклад тексту священної книги Тори, П'ятикнижжя Мойсеєвого, а саме – книга «Числа».

Сформовані для книги «Числа» 36 числові вектори, значеннями яких є числова характеристика вузлів *PageRank*, що обчислена для кожної сформованої мережі G_i $(i = \overline{1,36})$, представлені у таблиці 3.3 (терміни відповідають вузлам мережі).

Таблиця 3.3

Векторне представлення за допомогою числової характеристики вузлів *PageRank* сукупностей текстів, сформованих із розділів книги «Числа»

	1	**2**	**3**	**4**	…	**33**	**34**	**35**	**36**
lord	0.0067	0.0072	0.0114	0.0103	**…**	0.0265	0.0258	0.0252	0.0252
son	0.03	0.05	0.0395	0.034	**…**	0.0225	0.0231	0.0225	0.023
family	0.0319	0.0197	0.0279	0.0259	**…**	0.0234	0.0228	0.0222	0.0224
child	0.0215	0.0328	0.0289	0.0222	**…**	0.0216	0.0203	0.0201	0.0204
israel	0.013	0.0109	0.0167	0.0122	**…**	0.016	0.0151	0.0149	0.0152

child~of~israel	0.0162	0.0108	0.0171	0.0128	…	0.015	0.0144	0.0144	0.0151
offering	0	0	0	0.0022	…	0.0147	0.0144	0.014	0.0137
land	0.0045	0.0019	0.0019	0.0012	…	0.0122	0.013	0.0134	0.0135
mose	0.0065	0.0078	0.0103	0.0133	…	0.0125	0.0121	0.0121	0.0121
man	0.0238	0.0086	0.0111	0.0081	…	0.0107	0.0104	0.0104	0.0102
year	0.0378	0.0117	0.0118	0.0096	…	0.0107	0.0102	0.0099	0.0096
prince	0.0048	0.0272	0.0258	0.0188	…	0.0082	0.0084	0.0082	0.0081
tent	0.007	0.0038	0.0068	0.0129	…	0.0082	0.0079	0.0076	0.0075
father	0.0334	0.0278	0.0203	0.0167	…	0.0077	0.0073	0.0071	0.0075
tribe	0.0196	0.0315	0.0147	0.0124	…	0.006	0.0062	0.0061	0.0069
congregation	0.0213	0.0114	0.0082	0.0058	…	0.0069	0.0067	0.007	0.0068
people	0	0	0	0	…	0.0074	0.0072	0.0069	0.0068
tent~of~meeting	0.007	0.0088	0.0089	0.018	…	0.0071	0.0068	0.0066	0.0065
shekel	0	0	0.0046	0.0031	…	0.0068	0.0067	0.0065	0.0064

house	0.0428	0.0268	0.0192	0.0186	…	0.0068	0.0065	0.0063	0.0061
aaron	0.0185	0.0038	0.0077	0.0157	…	0.0062	0.0061	0.0059	0.0058
border	0	0	0	0	…	0.0026	0.0053	0.0058	0.0057
hand	0	0	0	0.0043	…	0.0051	0.005	0.0053	0.0053
priest	0	0	0.0092	0.0065	…	0.0045	0.0044	0.0053	0.0052
service	0	0	0.014	0.0199	…	0.0056	0.0054	0.0052	0.005

У відповідності з результатами, що представлені у таблиці 3.3, для кожної окремої послідовності числових значень *PageRank*, що відповідає певному терміну, було побудовано графік динаміки значень *PageRank*. Наприклад, для ключового терміна «lord», після застосування алгоритму побудови динамічної мережі термінів, графік динаміки *PageRank* цього терміна представлений на рисунку 3.3.

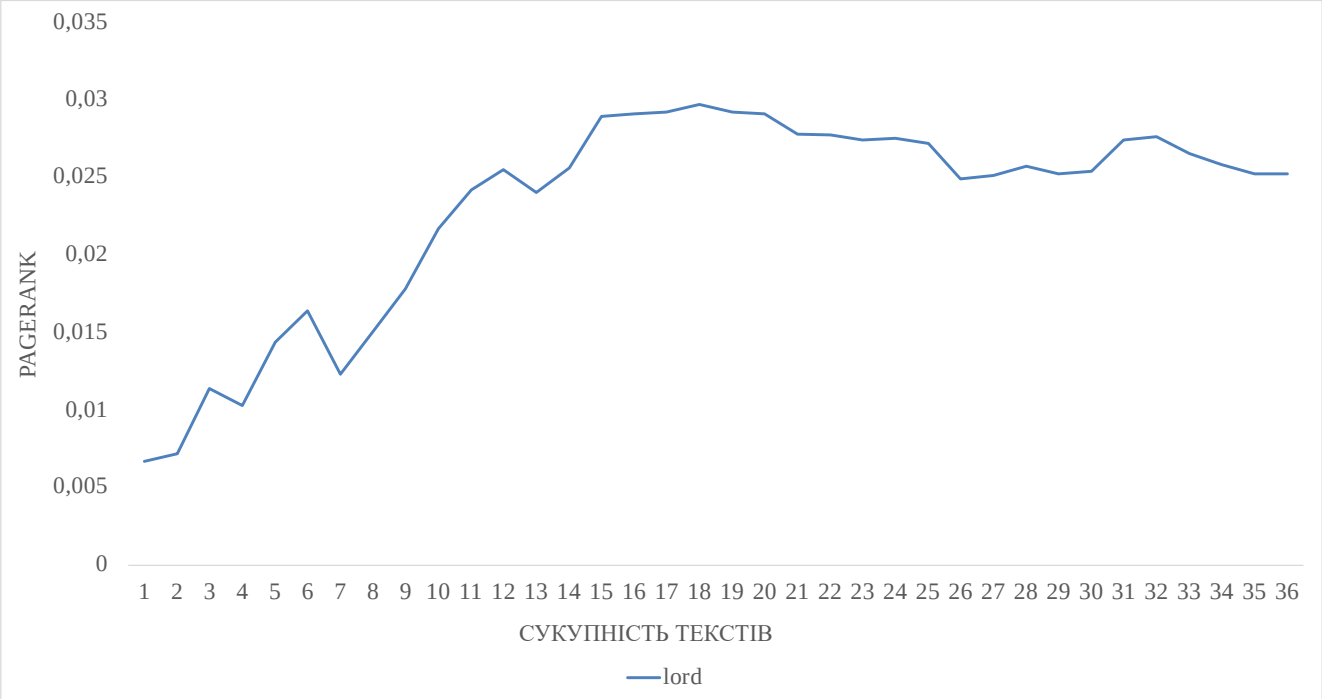

Рис. 3.3. Графік динаміки *PageRank* терміна «lord» для різної сукупності текстів T_i, сформованої з розділів книги «Числа»

Використовуючи результати, представлені у таблиці 3.3 було отримано зведений графік динаміки *PageRank* 10-ти найвагоміших ключових термінів – рисунок 3.4.

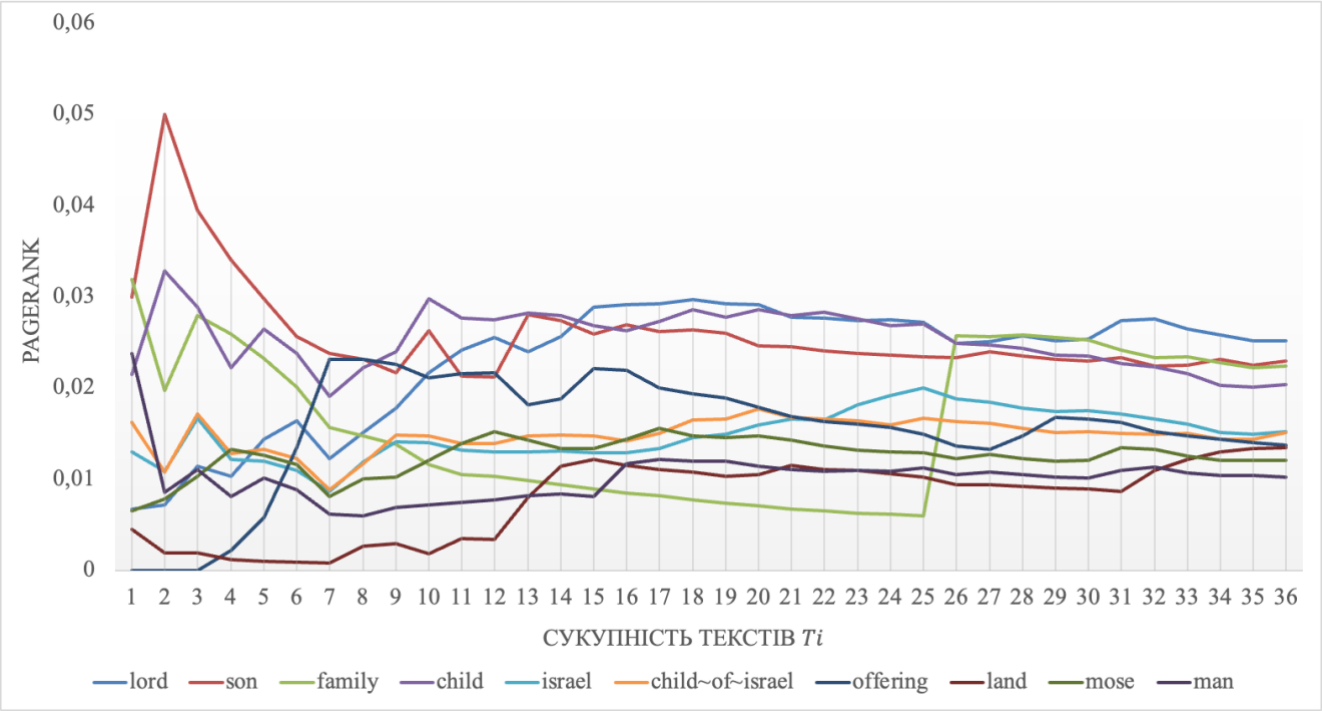

Рис. 3.4. Зведений графік динаміки *PageRank* 10-ти найвагоміших ключових термінів для різної сукупності текстів T_i, сформованої з розділів книги «Числа» в результаті застосування алгоритму побудови динамічної мережі термінів

На графіку, що представлений на рисунку 3.4 можна помітити, що вузли, які відповідають термінам «son», «family» та «child», підвищили значення показника *PageRank* у момент підвищення відповідних числових значень глобального показника важливості *GTF*. Для вузла, що відповідає терміну «son», таке підвищення числових значень *RageRank* спостерігається під час розширення сукупності текстів T_i 1-м та 2-м розділами книги «Числа». Для вузла, що відповідає терміну «family», підвищення значень *RageRank* відбувалося під час розширення сукупності текстів T_i 3-м розділом, та особливо стрімким було підвищення значень RageRank на 26-му кроці після розширення сукупності 26-м розділом книги, де

приріст зустрічання терміна «family» був найвищим впродовж всіх кроків роботи алгоритму побудови динамічної мережі термінів. Для терміна «child» підвищення значень *RageRank* корелює з підвищенням значень *GTF* цього терміна після розширення сукупності текстів 1-м розділом книги.

3.5.Методика визначення ступеня подібності текстових документів

Як було зазначено вище, прикладом онтології моделі, в якості якої можна представити тестові дані, та яка буде зручною для обробки комп'ютером, є направлена зважена мережа термінів (Directed Weighted Network of Terms – DWNT) або просто мережа термінів. Порівняння семантичних моделей представлення тексту, де вузлами такої мережі є ключові терміни (слова та словосполучення), а ребра – семантико-синтаксичні зв'язки між цими термінами DWNT, отриманих для різних текстів, дає змогу визначити семантичну близькість відповідних текстів.

В цій дисертаційній роботі представлена методика порівняння текстових документів та визначення ступеня подібності (розбіжності) між ними, що базується на побудові та порівнянні відповідних їм семантичних мереж [24, 26-28, 32]. Під час порівняння семантичних мереж, що відповідають текстовим документам, застосовується загальноприйнятий підхід, який полягає у наступному. Розглядається матриця $\boldsymbol{A - A'}$, яка є різницею матриць $\boldsymbol{A}$, $\boldsymbol{A'}$, що відповідають цим семантичним мережам і оцінюється її норма, як міра розбіжності. Норма матриці відображає порядок величини матричних елементів. У даному випадку рекомендується використовувати норму Фробеніуса $\|.\|_{\boldsymbol{F}}$, що дорівнює кореню квадратному із суми квадратів всіх елементів відповідної матриці [129]:

$$\|\boldsymbol{A - A'}\| = \sqrt{\sum_{\boldsymbol{i=1}}^{|\boldsymbol{A}|}\sum_{\boldsymbol{j=1}}^{|\boldsymbol{A}|}\left(\boldsymbol{a_{ij} - a'_{ij}}\right)^{\boldsymbol{2}}}. \qquad (3.4)$$

Звісно, що розмірність двох матриць, що порівнюються, має співпадати, тобто $|A|=|A'|$. У реальності склад термінів у різних семантичних матрицях

відрізняється. Тому мережі, що порівнюються, взаємно доповнюються термінами, що входять до їх загального складу.

3.6. Приклад апробації методики

Визначення ступеня подібності текстів було здійснено на прикладі біблійських текстів, які загальновідомі і перекладені майже на всі мови (зокрема, авторами досліджувались тексти івритом, китайською, англійською, російською і українською мовами). Для побудови мереж термінів й подальших досліджень був використаний український переклад тексту священної книги Тори, П'ятикнижжя Мойсеєвого, здійснений Іваном Огієнком [128] Загалом було опрацьовано всі п'ять книг - «Буття», «Вихід», «Левит», «Числа» та «Повторення закону».

В результаті опрацювання цих текстів було отримано онтологічні моделі у вигляді мережі із термінів, на Рис. 3.5 наведено фрагмент мережі термінів, що відповідає четвертій книзі «Числа», наведеній у відомому стенфордському перекладі українською мовою. Під час опрацювання «П'ятикнижжя Мойсеєвого», враховуючи специфіку священного письма, на етапі попередньої обробки текстів стандартний список стоп-слів корегувався: окремо формувався список слів-виключень, які не є стоп-словами та, насправді, є інформаційно-важливими; і навпаки, список стоп-слів доповнювався іншими словами, які не мають смислового навантаження в межах досліджуваного текстового документу.

Окремо опрацьовувались найбільш частотні слова-синоніми, яким в результаті присвоювалась єдина визначена лексема. Також у зв'язку з наявністю у текстах подібного стилю архаїзмів під час PoS-tagging деяким словам могли присвоюватись невірні теги, що потребувало ручного втручання.

Глобальність під час обчислення GTF визначалася в межах всієї книги, або в межах кожного окремого розділу залежно від того, для якого тексту будувалась мережа термінів – для всієї книги чи окремого розділу. Тому одні й ті ж терміни можуть мати різні значення GTF у межах окремого розділу та всього тексту, відповідно, що впливає на побудову графу горизонтальної видимості.

Щоб досягти незначної розрідженості матриць було також проведено видалення ребер, що мають одиничну вагу. Опісля також здійснювалось видалення вузлів, які не мають з'єднань. Такі вузли могли з'явитися, зокрема, і в результаті розрідження матриці.

Все вищесказане в результаті впливає на топологію мереж і призводить до наступних наслідків: у мережі термінів, що побудована для всієї книги, можуть бути вузли, яких не існує для кожного окремого розділу, та навпаки – мережа термінів для окремого розділу може містити вузли, яких немає у загальній мережі, що побудована для всього тексту.

Подальше порівняння за допомогою міри Фробеніуса матриць направлених зважених мереж, що отримані для різних текстів, дозволяє визначити семантичну близькість та ступінь подібності відповідних текстів.

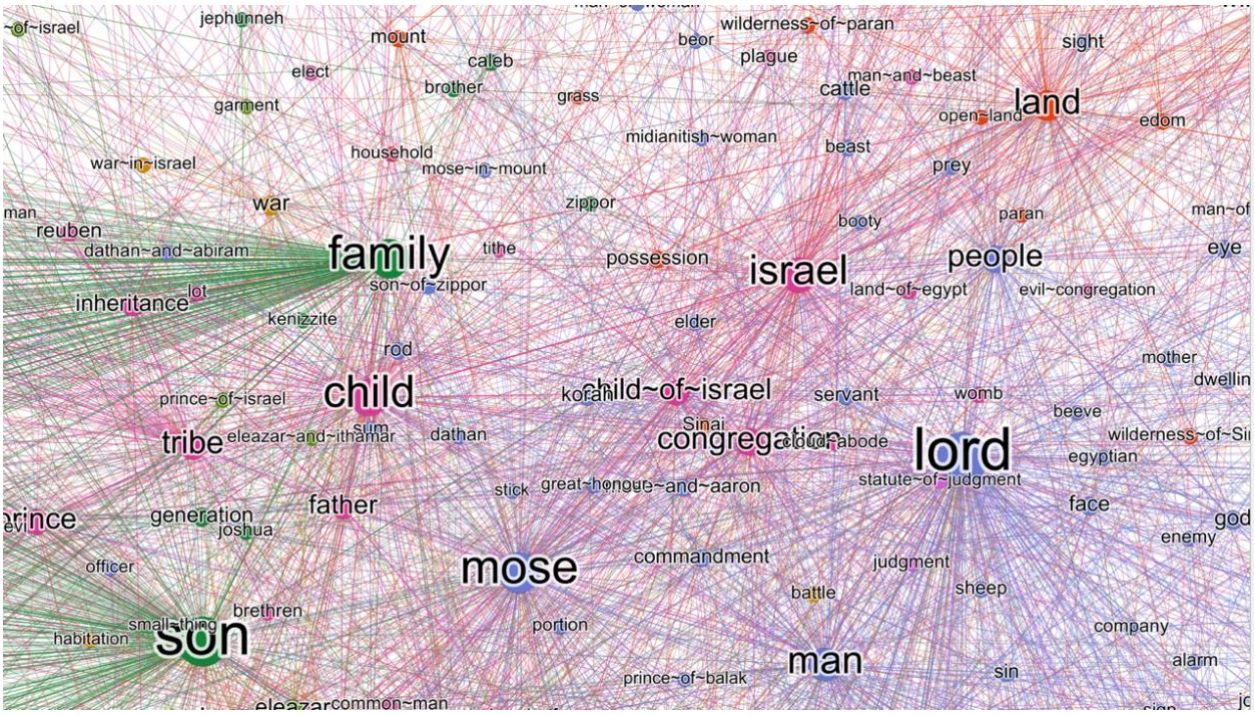

Рис. 3.5. Фрагмент мережі термінів, побудованих за книгою «Числа»

Книга «Числа» (четверта частина П'ятикнижжя Мойсея та Старого Заповіту) за своїм змістом найбільш наближена до правового документа, вона містить перепис ізраїльського народу за дорослими чоловіками, коли він перебував на Синайському півострові та на рівнині Моав та регламентує правила життя цього народу.

Перша частина книги життя охоплює 10 розділів. У них розповідається про останні дні перебування народу під Синаєм.

Друга частина (розд. 10-22) охоплює «40 років» перебування у пустелі.

У третій частині (розд. 22-36) описано події на землі Моава, зокрема пророцтва Валаама про благополуччя Ізраїля.

За всіми розділами цієї книги також були побудовані семантичні мережі (Рис. 3.6), які взаємно семантично порівнювались за ознакою збіжності за Фробеніусом.

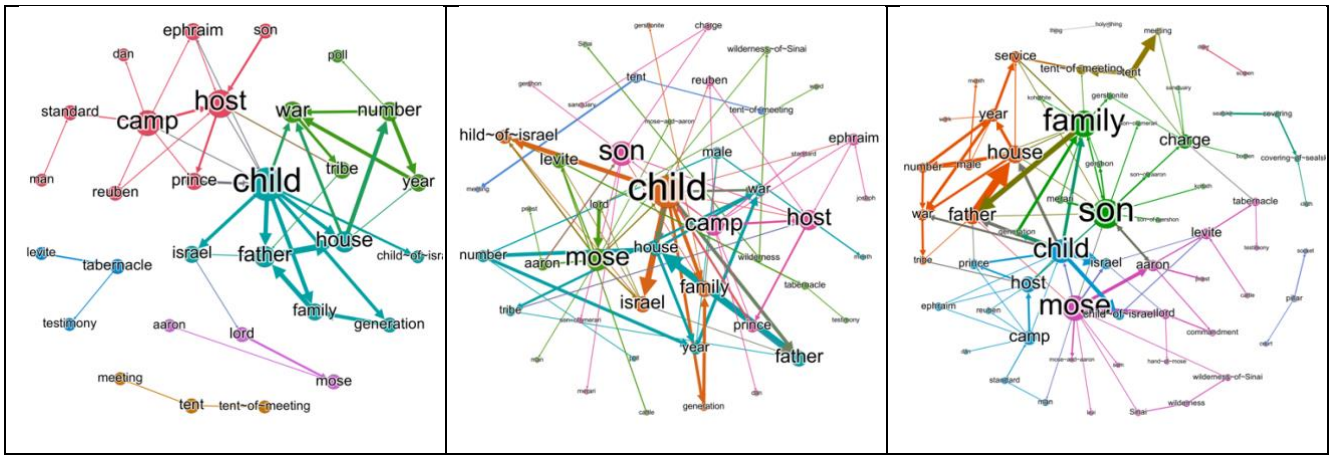

Рис. 3.6. Спрощені семантичні мережі, що відповідають окремим розділам книги.

На рис. 3.7 представлено графік визначення розбіжностей семантичних мереж за правилом порівняння матриць за Фробеніусом.

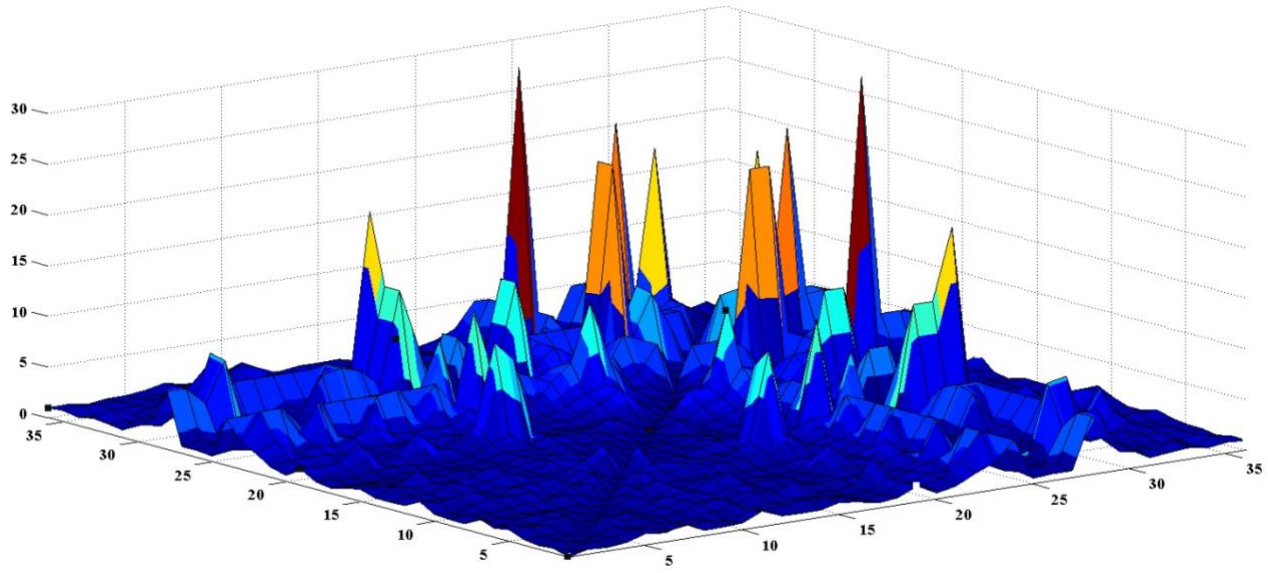

Рис. 3.7. Графік розбіжностей семантичних матриць, що відповідають окремим розділам книги «Числа».

Як можна побачити на графіку, найбільші значення розбіжностей відповідають третій частині, тобто розділам 22-36. Суть цієї аномалії можна знайти у дослідників Святого письма. Традиційно авторство книги приписується Мойсею, як авторові П'ятикнижжя. Разом з цим, описуються події, коли наступником Мойсея вже було обрано Ісуса Навина. Суто наративні фрагменти у цій частині книги переплітаються з юридичними приписами.

Тобто зміст книги «Числа» підтверджує наведену мережеву методику дослідження текстових документів щодо виявлення структурних і термінологічних розбіжностей. Саме книга «Числа» є самою близькою за змістом і структурою частиною Святого Письма до сучасних правових документів, що дозволяє обґрунтовано припустити, що наведена методика може застосовуватись і до таких документів, зокрема, при здійснені парламентського контролю [4, 10, 12, 21-23, 26].

Висновки до розділу 3

В результаті огляду було встановлено, що у прикладних дослідженнях зазвичай застосовують типові для мережевого аналізу характеристики вузлів мереж, найважливішими серед яких на цей час вважаються степінь вузла та показники, що відповідають алгоритмам HITS та PageRank.

В цьому розділі дисертаційної роботи також було запропоновано алгоритм побудови динамічної мережі термінів та за його допомоги досліджено динаміку вагових значень вузлів у мережі термінів. Використовуючи алгоритм побудови динамічної мережі термінів можна досліджувати динаміку окремих ключових термінів в результаті підвищення чи зниження їх глобальної частоти зустрічання у тексті шляхом додавання текстових документів, які насичені чи збагачені окремим визначеним терміном, у інформаційний потік. Такі термінологічні збагачення можуть бути штучними й викликані «інформаційними вкидами», пропагандою чи

спамом. Також вони можуть бути результатом навмисних, цілеспрямованих інформаційних атак – інформаційних операцій. Тож їх виявлення може бути здійснене шляхом аналізу динаміки ключових термінів, отриманої в результаті застосування алгоритму побудови динамічної мережі термінів.

У цьому розділі викладено методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж. Ця методика може стати основою побудови систем порівняння правових документів у рамках парламентського контролю. Також розглянуто алгоритм побудови семантичних мереж як одного із видів онтологій. Цей алгоритм також може застосовуватися в системах автоматичного реферування правової інформації з метою формування лаконічних інформаційно-насичених звітів, коротких анотацій або дайджестів. Пропонована методика може бути використана в процесі обробки запитів при проведенні інформаційного пошуку, надаючи можливість визначення ступеня подібності або відмінності структури та семантики текстів.

РОЗДІЛ 4. ТЕХНОЛОГІЧНІ ЗАСАДИ ФОРМУВАННЯ МЕРЕЖЕВОЇ МОДЕЛІ ПРЕДМЕТНОЇ ГАЛУЗІ

В цій дисертаційній роботі представлена цілісна технологічна схема виокремлення та формування ключових термінів тексту та формування мережевої моделі ключових термінів. В цьому розділі також пропонується лінгвостатистичний метод автоматичного екстрагування, дослідження динаміки і виявлення взаємозв'язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів.

Представлена модель середовища семантичного інформаційного пошуку та метод ранжування інформаційних джерел.

Наприкінці розділу розглядається методика використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій.

4.1.Технологічна схема екстрагування ключових термінів

У цьому розділі представлена цілісна технологічна схема виокремлення ключових слів та формування термінів із текстів інформаційних повідомлень [30].

На першому етапі технологічної схеми здійснюється послідовна обробка отриманого на вхід природомовного тексту за допомогою функцій Pipeline відповідних бібліотек pymorphy2 [110], Stanza [111] або NLTK[112]. Далі здійснюється формування та виокремлення ключових слів, біграм та триграм за визначеними шаблонами, та їх подальше статистичне зважування за частотою появи у тексті. Наприкінці здійснюється вивід найбільш частотних термінів у вигляді списку слів, біграм та триграм поданих у нормальній словниковій формі (для триграм третій елемент не нормалізується) та їх найчастотніших ненормалізованих вхідних форм в тому вигляді, в якому вони зустрічаються у вхідному тексті повідомлення чи текстової публікації.

Для попередньої обробки текстових даних застосовуються описані у попередніх розділах найпоширеніші прийоми, що включають автоматичну сегментацію на окремі речення та подальшу токенізацію тексту – сегментації

вхідного тексту на елементарні одиниці (токени, лексеми). В межах кожного речення після токенізації здійснюється розмічування частин мови (Part-of-Speech tagging) [17, 19, 20], що полягає у віднесенні слова в тексті до певної частини мови й присвоєні йому відповідного тега. PoS tagging дозволяє розділити слова або токени, які можуть мати декілька тегів. Додатково здійснюється лемматизація окремих розмічених лексем з метою отримати їх канонічні, словникові форми – леми. Цей крок дозволяє додатково згрупувати різні форми одного й того слова, щоб їх можна було проаналізувати як єдиний елемент.

Серед розмічених слів виокремлювались уніграми, які належать до іменників. Для побудови словосполучень використовуються визначені шаблони.

Далі здійснюється видалення одиничних стоп-слів (окремих артиклів, прийменників, сполучників, деяких дієслів, прислівників та займенників), які не несуть ніякого інформативного навантаження.

Кінцеве статистичне зважування та ранжування виокремлених термінів здійснювалось шляхом обчислення їх загальної частоти появи у тексті окремого інформаціного повідомлення [3-8].

Після виокремлення ключових термінів та їх статистичного зважування, в межах кожного окремого речення формується послідовність, де слова розташовуються у тому порядку, в якому вони зустрічаються у реченні тексту, а словосполучення з більшою кількістю слів розташовуються перед словосполученнями та словами, які є їхньою частиною. Для кожного речення тексту формується окрема послідовність термінів. Після цього здійснюється формування ненаправлених зв’язків між цими термінами.

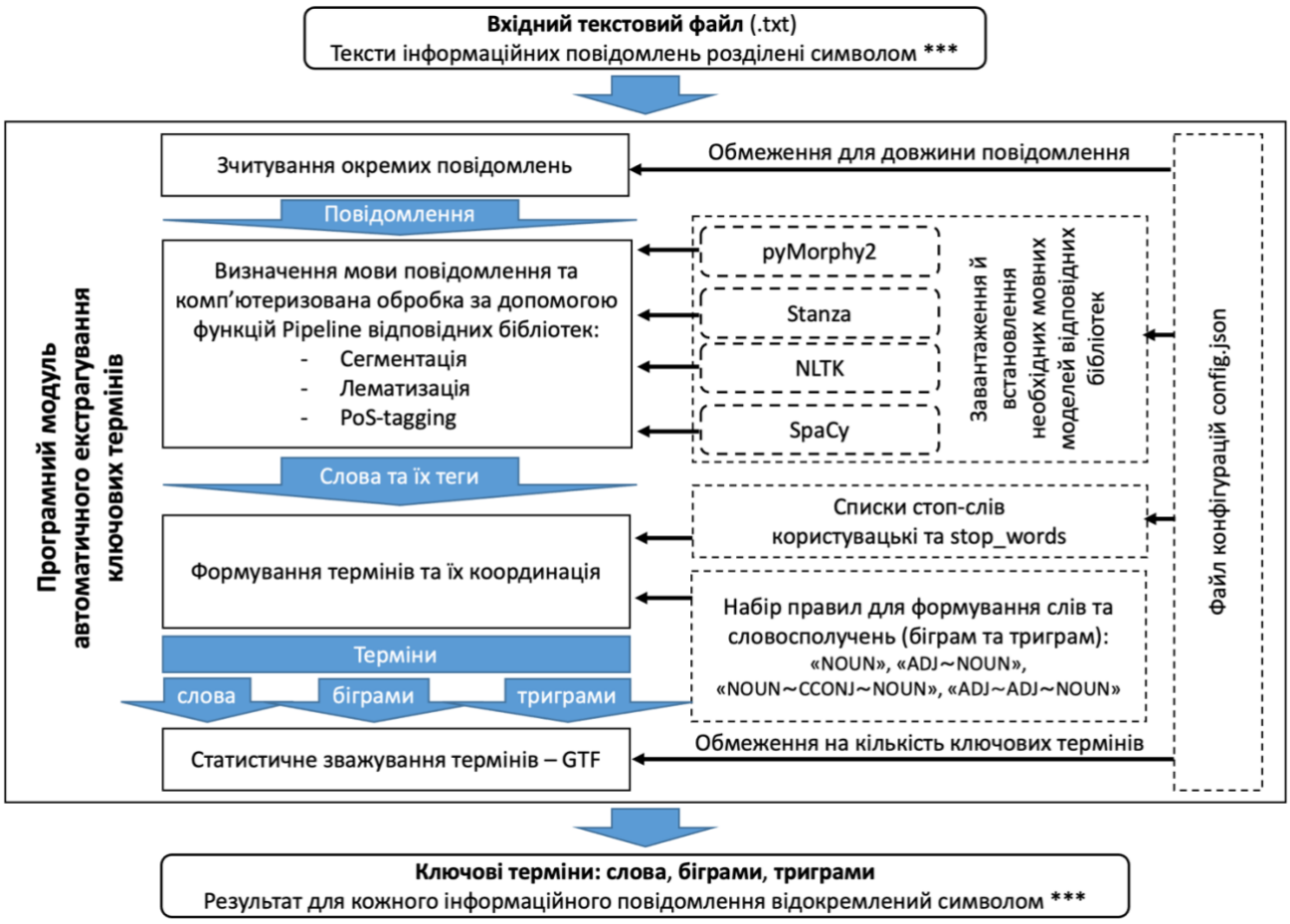

Рис. 4.1 Технологічна схема формування ключових термінів

Для побудови ненаправленої мережі для послідовності слів та словосполучень, у відповідність яким поставлені вагові значення в цій дисертаційній роботі використовується метод побудови графів видимості для ключових термінів (окремих уніграм/слів, біграм та триграм), зокрема – алгоритм побудови графа горизонтальної видимості (Horizontal Visibility Graph algorithm – HVG) [116, 117, 118].

4.2. Екстрагування і виявлення взаємозв'язків фразеологізмів в інформаційних потоках

В роботі пропонується лінгвостатистичний метод автоматичного екстрагування і виявлення взаємозв'язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності

фразеологізмів [31]. Засоби виявлення сталих словосполучень будуються на основі концепцій машинного навчання, лінгвістичного аналізу і статистичних розрахунків. У подальшому за допомогою сучасних методів аналізу мереж досліджуються взаємозв'язки фразеологізмів, визначаються їх окремі кластери, які ймовірно відповідають наративам. Запропоновано форму візуального відображення інформаційного потоку в розрізі фразеологізмів і дат, що є прямокутною таблицею (Phraseme Diagram), комірки якої заповнені чисельними значеннями, що відповідають частотам фразеологізмів у розрізі дат. Розглянутий підхід може застосовуватись для вирішення питань аналізу та візуалізації розподілу наративів для будь-яких відібраних інформаційних масивів у розрізі питань, що цікавлять дослідника та мають досить значні часові рамки.

Відомо, що наратив (від лат. narrare – розповідати, пояснювати) – сукупність пов'язаних між собою реальних чи вигаданих подій, фактів або вражень, які складають оповідний текст. Головна функція наративу – це репрезентація одиничної події або численних подій. Наратив організовує людський досвід у складений зі знаків текст [130, 131].

У різні часи в інформаційних потоках відображаються свої наративи, що динамічно змінюються. З ретроспективного погляду дослідження їх змін може пояснити багато речей, наприклад, умисний вплив окремих кіл на суспільну думку, що, наприклад, на прямую впливає на загальну безпеку держави. Зловмисник може впроваджувати різні наративи у суспільний дискурс, впливаючи на суспільну свідомість. Аналіз динаміки наративів дозволяє прогнозувати суспільні явища, процеси. Досліджуючи штучно створені наративи можна виявити ймовірні приховані цілі супротивника. Дослідження старих наративів дозволяє перейти від ретроспективного аналізу до прогнозу. Визначення нових наративів може сприяти оперативній розвідці, виявленню ворожих намірів. Тому задача дослідження наративів особливо актуальна в умовах інформаційних або гібридних війн.

Безпосереднє дослідження наративів в інформаційних, їх автоматичне виявлення, це складна задача, яку можна розкласти на окремі часткові задачі, серед яких можна виділити задачу автоматичного виявлення окремих фразеологізмів,

сукупність яких утворює наратив. Фразеологізми (фразеологічні звороти) прийнято називати усталені словосполучення, які сприймаються як єдине ціле і вживаються носіями мови в усталеному оформленні. До фразеологізмів, серед іншого, відносяться фразеологічні сполучення і фразеологічні вирази (стійкі фразеологічні обороти).

Вочевидь, взаємозв'язок фразеологізмів (а точніше, фразеологічних виразів), дозволяє виявити найбільш зв'язані за змістом, тобто реконструювати наратив.

Задача виявлення фразеологізмів відноситься до класу задач розпізнавання текстових образів (Text Pattern Recognition) [132], близька за змістом до задач виявлення із текстових документів понять (Concept Recognition) [133], подій (Event Recognition) [134], знаходження мемів (коротких лексичних позначень явищ, процесів). Для вирішення можна застосувати такі методи:

1. Шаблонний (Template Methods), коли необхідні фрагменти текстів знаходяться за збігом деяких фрагментів тексту. Шаблони формуються аналітиками. Цей метод відноситься до методу навчання із вчителем

2. Традиційні методи машинного навчання (Machine Learning), також методи навчання із вчителем, коли необхідні частини документів знаходяться за допомогою алгоритмів машинного навчання, які застосовують заздалегідь сформовані моделі. Ці моделі є комп'ютерним узагальненням корпусів (розмічених масивів) текстових документів (або фрагментів), для яких експерти вже зробили висновок щодо цільової корисності.

3. Лінгвостатистичні методи, що базуються на автоматичному виявленні сталих словосполучень і емпіричних методів розрахунку міри їхнього впливу, ваги. Ці методи відносяться до навчання без вчителя.

Шаблонні методи можуть застосовувати цілі стійки фрагменти текстів, або множини слів і словосполучень. Ці методи можуть легко інтегруватися із вже існуючими пошуковими системами, і це, звичайно, перевага. Але принциповий недолік цих методів, це практична неможливість виявлення нових фразеологізмів. Лише після того, як вони будуть знайдені експертами, вони зможуть

ідентифікуватись в нових документах. Ще один недолік цих методів – невисока повнота і точність.

Методи машинного навчання можуть бути переважними лише при наявності великих за обсягів і точних за змістом моделей. Такі методи охоплюють десятки алгоритмів – від наївних байєсовських (Naive Bayes Algorithm), найбільш швидких, до багатошарових нейронних мереж (Deep Learning). Для створення моделей експертами мають бути відібрані великі набори (до рівня Big Data) релевантних інформаційній потребі документів з визначеними фрагментами, які є наративами. І ця робота вимагає значних людських і часових ресурсів, що є визначним недоліком. До переваг таких моделей при умові їх якісного налаштування – повнота і точність.

Лінгвостатистичні методи також можна розглядати як методи машинного навчання, але без вчителя. Передбачається, що фразеологізми мають виявлятися автоматично як сукупність найбільш частотних словосполучень, які мають найбільшу вагу, що визначається за правилами, які одного разу визначаються експертами, виходячи із їхнього досвіду. Виходячи з цього, такі методи дозволяють виявляти нові фразеологізми і не вимагають значних зусиль на ручну попередню розмітку і аналіз ретроспективного фонду. Повнота отриманих результатів залежить лише від налаштувань довжини словосполучень, але питання точності при обмеженні цієї довжини вимагає окремих досліджень.

В межах цієї роботи пропонується до розгляду лінгвостатистичний метод. Цей метод передбачає застосування масивів для дослідження і навчання достатньо великого обсягу (метод було апробовано на десятках тисяч релевантних документів) відсортованого за датами публікацій. Сутність методу полягає у виконанні таких технологічних операцій, як експертне створення запиту до наявних інформаційно-пошукових систем [135, 136], що відповідає об’єкту зацікавленості. В результаті опрацювання цих запитів створюються великі за обсягом масиви релевантних документів, в яких за допомогою спеціальних алгоритмів мають бути визначені необхідні фрагменти. На базі відібраних масивів (різних мовних версій) визначаються сталі словосполучення, що відносяться до різних періодів часу. Засоби виявлення сталих словосполучень будуються на основі

концепцій машинного навчання, лінгвістичного аналізу і статистичних розрахунків [8, 137]. У подальшому за допомогою сучасних методів аналізу мереж досліджуються взаємозв’язки фразеологізмів, визначаються їх окремі кластери, які ймовірно відповідають наративам.

У цій роботі запропонована форма візуального відображення інформаційного потоку в розрізі фразеологізмів і дат, що є прямокутною таблицею, будемо називати її Ph-Di діаграмою (Phraseme Diagram), комірки якої заповнені кількістю документів, що відповідають обраному фразеологізму в розрізі дат. Тобто, стовпцям цієї таблиці відповідають дати, а рядкам – фразеологізми, які можна розглядати як своєрідні змістовні фільтри інформаційного потоку.

Візуально запропонована Ph-Di діаграма є таблицею, комірки якої зафарбовані відтінками колірної гами, залежно від значень обсягів публікації за вибраним об'єктом (фразеологізмом) у відповідний день (велике значення відповідає більш світлому відтінку). Запропоновані діаграми для відносно невеликої кількості рядків-фразеологізмів (кілька десятків) дозволяють без додаткової обробки виявляти групи найбільш пов'язаних за датами та інтенсивністю публікацій об'єктів візуально. Для великої кількості об'єктів у процесі побудови діаграми пропонується її кластеризація, за результатами якої здійснюється перестановка рядків (перегрупування фразеологізмів). Для кластеризації пропонується сформувати мережу взаємозв’язку фразеологізмів і виявити групи найбільш зв’язаних між собою і віддалених від інших(кліки). Передбачається, що отримані кластери із тісно пов'язаних фразеологізмів і відповідатимуть наративам.

Запропонований в цій роботі метод екстрагування, дослідження динаміки і виявлення взаємозв’язків фразеологізмів в інформаційних потоках передбачає виконання низки кроків, а саме:

Крок 1. Формування стартових запитів (шаблонів для вибору текстів), що відповідають загальній тематиці. Як приклад, можуть застосовуватись вимоги пошуку в новинах із Інтернету фрагментів тексту, де є посилання на Україну. У цьому випадку запит має вигляд:

україн & (влада | армія | народ | суспільство | населення)

Крок 2. В результаті опрацювання створених на першому кроці запитів створюються великі за обсягом масиви релевантних документів, в яких за допомогою спеціальних алгоритмів мають бути визначені необхідні фрагменти. Фрагменти являють собою речення, які відповідають запиту, крім того для кожного такого речення до вихідного файлу додаються також сусідні речення.

Крок 3. Сутність третього кроку - витяг найважливіших окремих слів (уніграм), і словосполучень із файлів, отриманих на Кроці 2. Для цього запускається програмний модуль виокремлення ключових термінів (слів і словосполучень) з тематичних інформаційних потоків для подальшого пошуку фразеологізмів termsExtractor. Цей модуль призначений для попередньої обробки природомовних текстових даних тематичних інформаційних потоків, що включає токенізацію тексту та видалення стоп-слів, і подальше виокремлення ключових слів і словосполучень за допомогою застосування більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging) та кінцевого статистичного зважування та ранжування термінів за частотою їх появи для подальшого виявлення фразеологізмів. Для реалізації цього модуля мають застосовуватись спеціальні бібліотеки типу pyMorphy2, NLTK, SpaCy, Stanza та fastText (для ідентифікації мов).

Крок 4. На 4-у кроці має здійснюватись вибір найбільш частотних слів за окремими датами (саме вони у запропонованій моделі будемо вважати фразеологізмами), а після цього, вибір набільш частотних фразеологізмів, що відповідають всім датам. Саме ці фразеологізми будуть надалі використовуватись при візуалізації та при побудові графів взаємозв'язку фразеологізмів.

Крок 5. На 5-му кроці здійснюється відображення Ph-Di діаграми в описаному вище вигляді. Приклад такого відображення приведено на Рис. 4.2. На цій діаграмі яскраві горизонтальні риски (висока частота окремого фразеологізму за деякий період часу) можуть у явному вигляді підказувати користувачу щодо трендів суспільної думки.

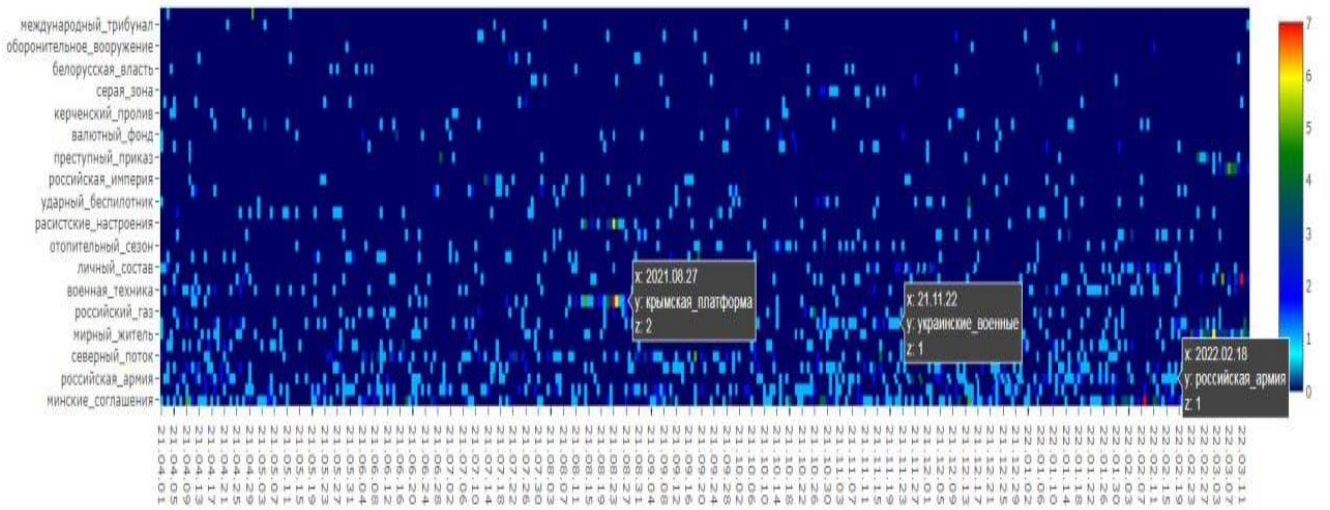

Рис. 4.2. Приклад візуалізації Ph-Di діаграми: горизонтальна вісь – дати, вертикальна вісь – номери фразеологізмів, от сортовані за частотою появи, яскравість точок відповідає абсолютній частоті появи фразеологізму за визначену дату

Крок 6. На 6-му кроці розраховується матриця взаємозв'язку фразеологізмів. Окремі фразеологізми вважаються зв'язаними, якщо вони одночасно входять до одного й того ж фрагмента тексту. Вага такого зв'язку між двома фразеологізмами відповідає кількості фрагментів тексту документів, що відповідають одночасно двом фразеологізмам. На основі сформованої матриці взаємозв'язку здійснюється відображення відповідного графу – мережі фразеологізмів, приклад якої наведено на Рис. 4.3.

Крок 7. На останньому, 7-му кроці здійснюється кластеризація мережі взаємозв'язку фразеологізмів. Кластеризація може здійснюватись за допомогою різних алгоритмів, зокрема на основі розрахунку модулярності мережі. На основі проведеної кластеризації шляхом експертного аналізу може здійснюватись визначення наративів, як узагальнень фразеологізмів, що спільно входять до однакових кластерів. В наведеній вище таблиці можна виділити, наприклад, такі групи наративів, як «Північний потік»: («Северный поток» и «Российский газ»); «Війна в Україні»: («Российская армія», «Боевые действия», «Военная операція», «Военная техника»); «Ситуація в Криму»: («Кречeнский пролив», «Крымский

парламент»). Як можна бачити, мережі взаємозв’язків фразеологізмів забезпечує відносно високу повноту, для забезпечення високої точності необхідне залучення експертів.

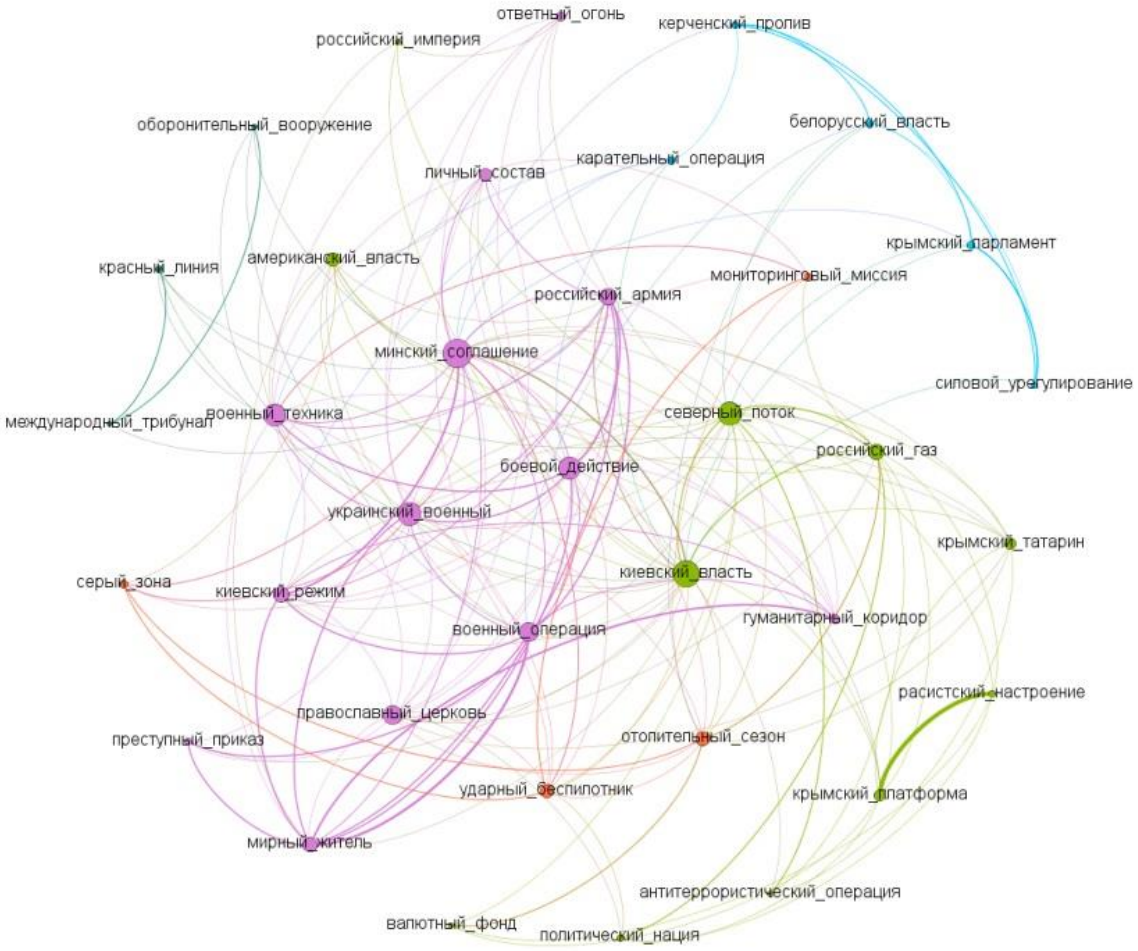

Рис. 4.3. Приклад візуалізації мережі взаємозв’язків фразеологізмів

Таким чином, для реалізації запропонованого методу необхідно було виконати наступні роботи:

Створити набір стартових запитів до наявних інформаційно-пошукових систем, які охоплюють великі обсяги інформації (ключових слів, сталих фрагментів текстів і операторів, що їх поєднують).

Розробити програмні застосунки витягу необхідних фрагментів із вибраних за допомогою інформаційно-пошукових систем документів.

Розробити програмне забезпечення виявлення сталих словосполучень на базі моделей машинного навчання (із застосуванням наявних вільних програмних бібліотек обробки природної мови).

Розробити/адаптувати програмне забезпечення формування мереж взаємозв’язку фразеологізмів, їх візуалізації, кластерного аналізу для подальшого виділення наративів.

Розробити інтерфейс користувача системи.

Запропонований лінгвостатистичний підхід поєднує переваги обох методів машинного навчання з вчителем. Представлений підхід до аналізу інформаційних потоків з метою виявлення наративів як сукупності фразеологізмів носить об'єктний характер, який, у свою чергу, представляється як суттєва складова методологічної бази прогнозного аналізу.

Метод він не вимагає великих людських і часових витрат. Саме ці витрати були вже внесені, коли протягом багатьох років формувались ретроспективні бази даних пошукових систем.

В результаті проведених експериментів є підстави припустити, що використання таких засобів візуалізації, як Ph-Di діаграми, дозволяє «розкладати» вихідні часові ряди відповідно до складу та особливостей фразеологізмів, виявляти активність публікацій, що відповідають певним наративам, виявляти взаємозв'язки фразеологізмів, визначати деталі динаміки народження в інформаційному потоці нових фразеологізмів.

Розглянутий підхід може застосовуватись для вирішення питань аналізу та візуалізації розподілу наративів для будь-яких відібраних інформаційних масивів у розрізі питань, що цікавлять дослідника та охоплюють значні часові рамки.

Треба підкреслити, що метод може застосовуватись лише тоді, коли у наявності є ретроспективний фонд, який дозволить створити базу даних статистичного машинного навчання необхідних обсягів.

4.3.Модель середовища інформаційного пошуку

В цій дисертаційній роботі запропонована модель середовища інформаційного пошуку [31], що базується на використанні методики побудови направленої зваженої мережі термінів (DWNT).

В процесі інформаційного пошуку на основі методики побудови направленої зваженої мережі термінів (DWNT) здійснюється наступне:

1. Кожному текстовому документу із масиву документів (бази даних) ставиться у відповідність спрощена семантична мережа, побудована за алгоритмом DWNT, запропонованим та розробленим у цій дисертаційній роботі [23, 25].

2. Кожна мережа представляється у вигляді пар зв'язаних вузлів – пар термінів.

3. Індексна система являє собою файл (таблицю), який для кожного документа містить як окремі записи (рядки) пари термінів і зв'язаний з ними ідентифікатор документа.

Як запит до системи може розглядатися деякий первинний документ, який спочатку задовольняє інформаційній потребі користувача. Якщо такого документа користувач не має у розпорядженні, він сам може змоделювати такий документ, записавши бажану інформацію природною мовою.

Пошук здійснюється таким чином:

1. За первинним документом формується мережа DWTN, з якої вибираються всі зв'язані пари термінів, що відповідають зв'язаним вузлам мережі.

2. Формується запит, в якому приймають участь всі пари термінів.

3. Користувачеві виводяться всі документи із бази даних, впорядковані за ознакою кількості співпадіння пар термінів в документах бази даних з парами термінів, сформованих в п.1.

Зазначимо, що кількість документів, що виводяться, може обмежуватись деяким порогом, може вимагатись співпадіння N або більшої кількості пар термінів, наприклад, для N=5.

4.4.Модель ранжування інформаційних джерел

У цій дисертаційній роботі запропонована нова модель ранжування як окремих документів, так і джерел інформації, що стосуються визначеної у інформаційному запиті проблемної галузі [33].

Сутність моделі полягає у тому, що користувач формує широкий інформаційний запит за тематикою тією мовою, що притаманна традиційним інформаційно-пошуковим системам. Якщо у результаті обробки запиту отримується масив інформації, що релевантний запиту, то за цим масивом формується його відповідна семантична мережа DWNT за правилами, описаними вище [19, 23, 25, 31]. Нагадаємо, вузлам цієї мережі відповідають ключові терміни або їх сталі сполучення [17, 18, 20, 31], а ребрам – зв'язки між ними, що визначаються правилами побудови DWTN. Ця мережа в рамках моделі розглядається як еталонна, з нею можуть порівнюватися мережі DWTN [26, 27], що формуються як за окремими документами, так і за сукупностями документів, що відносяться до окремих інформаційних джерел.

Порівняння будь-яких двох мереж N1 і N2 в рамках моделі пропонується здійснювати за таким алгоритмом [33]:

Крок 1. Обираємо два пороги t1 і t2. Якщо вага будь-якого ребра мережі N1 (N2) менша за t1 (t2), то вважаємо що відповідного зв'язку не існує, інакше він існує і його вага дорівнює 1. Значення t1 і t2 обираються експертно, наприклад, виходячи із міркувань необхідної щільності мережі, що розглядається, або із умов зв'язності цієї ж мережі.

Крок 2. Встановлюємо значення розбіжності між мережами N1 і N2, як P=0.

Крок 3. Для кожної пари термінів w1 і w2, що входять до N (перетину множин вузлів мереж N1 і N2), перевіряється, чи є ребро між ними у двох мережах одночасно. Якщо у одній мережі ребро є, а у іншій немає (зв'язок із вагою 0), то розбіжність P збільшується на одиницю.

Крок 4. Обчислене значення P нормується, з нього береться корінь квадратний і отримане значення ділиться на розмір (потужність) множини N.

Фактично за цим алгоритмом розраховується міра Фробеніуса матриці [26, 27], яка являє собою різницю між матрицями мереж N1 і N2, якщо в них враховуються лише вузли, що відповідають спільним термінам.

Якщо виникає необхідність інкрементного доповнення мережі DWNT, у випадку, коли до відповідного їй масиву документів додається новий документ, необхідно виконати такі кроки:

1. Поєднуються вузли. До первинної мережі DWNT додаються нові вузи мережі DWNT, яка відповідає документу, що додається.

2. Складаються значення зв'язків. Вага зв'язків первинної мережі DWNT збільшуються на значення ваги мережі, що відповідає новому документу. Якщо у первинній мережі зв'язку не було, а у новій мережі він є, то зв'язок додається в первинну мережу.

3. Напрямок зв'язків. Розглядається декілька випадків. а) якщо у первинній мережі зв'язків не було, а в новій є, в об'єднаній мережі встановлюються напрямки нової мережі. б) якщо у первинній мережі зв'язки були, вони залишаються такими ж самими. Вочевидь, напрямки для об'єднаної мережі необхідно періодично перераховувати на основі аналізу всього об'єднаного масиву документів.

4.5.Апробація моделі ранжування інформаційних джерел

Продемонструємо рейтингування мережевих джерел інформації відносно тематики, що відповідає розповсюдженню штаму вірусу COVID-19 (Delta Variant) і процесам інфляції в економіці.

Для цієї тематики складено широкий запит до системи контент-моніторингу:

"Delta Variant" & Inflation

За перший тиждень вересня 2021 року за цим запитом отримано близько 400 документів із понад 100 інформаційних джерел. Було розглянуто 6 джерел, яким відповідала найбільша кількість документів. Після цього були побудовані відповідні семантичні мережі. На рис. 4.4, як приклад, наведена семантична

мережа, що відповідає джерелу «Independent», а на рис. 4.5. – фрагмент еталонної мережі (позначеної в таблиці 4.1, як мережа джерела All).

У таблиці 4.1 наведено значення розбіжності, що розраховувалися за наведеним алгоритмом. Ці значення можуть виступати як основа рейтингування інформаційних джерел у відповідності із основною вибраною тематикою.

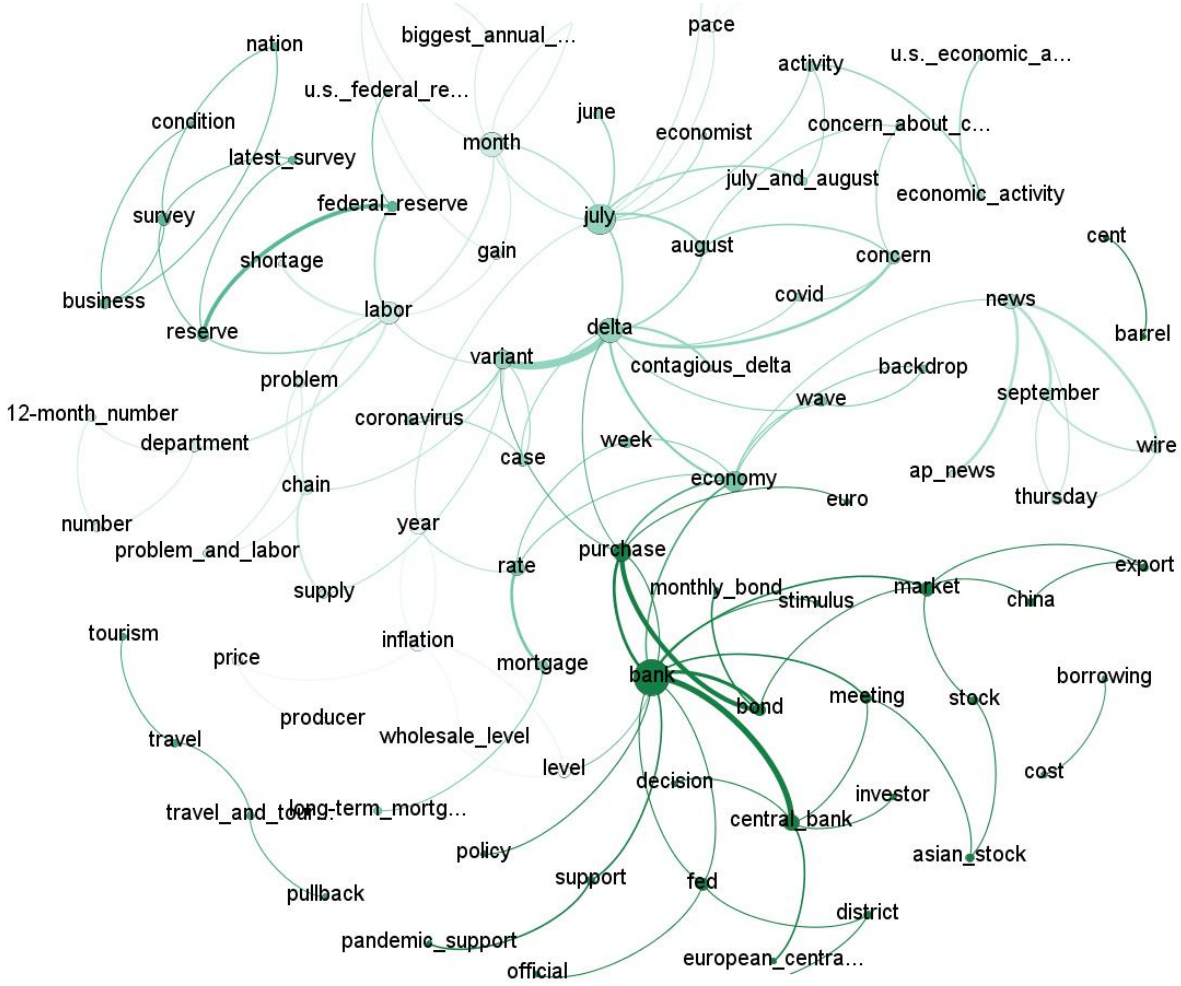

Рис. 4.4. Семантична мережа, що відповідає джерелу Independent.co.uk

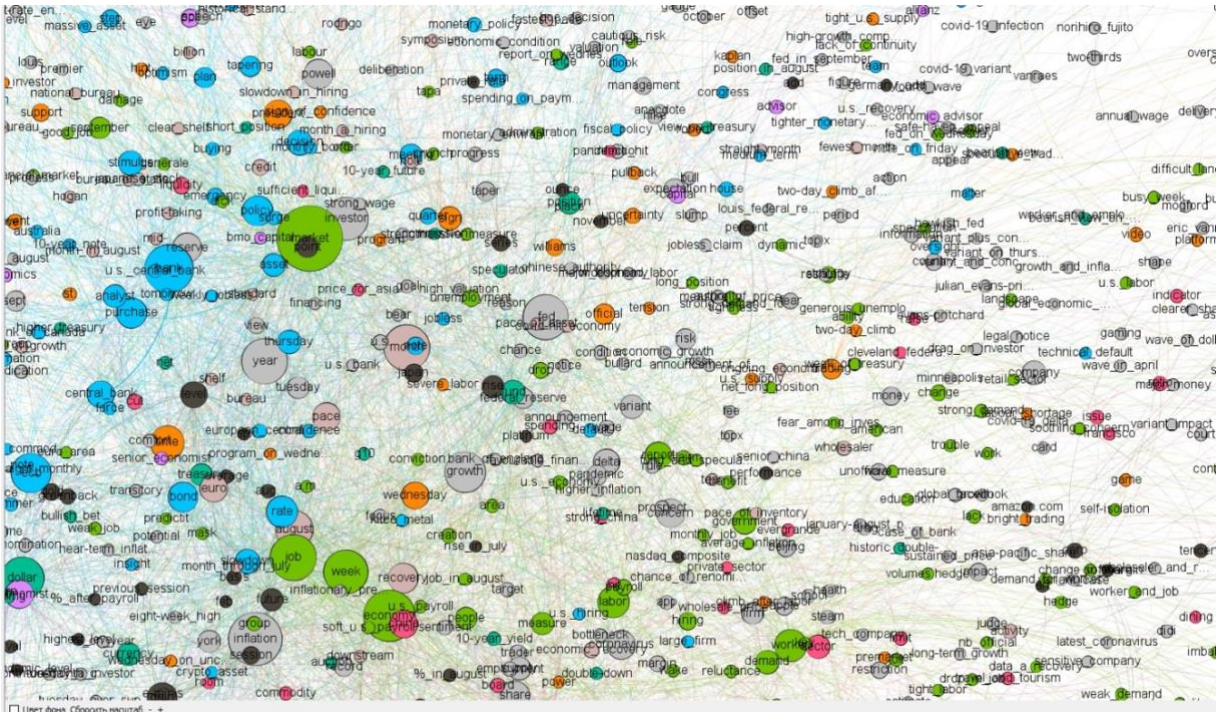

Рис. 4.5 Фрагмент еталонної мережі, що відповідає запиту “Delta Variant” & Inflation

Таблиця 4.1

Таблиця розбіжності між мережами, що відповідають тематиці і джерелам

Джерело	Адреса	Кількість знайдених	Розбіжність із еталоном
All	-	400	0
Associated Press	https://apnews.com	5	0,385
Seattle Times	https://www.seattletimes.com	5	0,263
Economic Times	https://economictimes.indiatimes.com	5	0,229
Independent	https://www.independent.co.uk	5	0,387
International Business Times	https://www.ibtimes.com	10	0,155
Reuters	https://www.reuters.com	70	0,033

Таким чином, у даному прикладі формується рейтинг джерел, щодо відповідності тематики інформаційного запиту на основі порівняння семантичних мереж:

1. Reuters (0,033)
2. International Business Times (0,155)
3. Economic Times (0,229)
4. Seattle Times (0,263)
5. Associated Press (0,385)
6. Independent (0,387)

4.6.Методика використання DWNT для формування БЗ СППР під час розпізнавання ІО

У цьому розділі запропоновано методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень (БЗ СППР) під час розпізнавання інформаційних операцій (ІО) [13, 14, 18].

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

Далі представлені послідовні етапи, що передбачають попередню комп’ютеризовану обробку текстового корпусу, такі як: токенізація (лексичний аналіз), лематизація, вилучення стоп-слів, стемінг та статистичне зважування й виокремлення ключових термінів. В якості функції, яка ставить у відповідність кожному терміну число, використовується запропонований у цій дисертаційній роботі статистичний показник важливості терміна GTF. Присвоївши кожному терміну число у порядку появи їх у тексті й виокремивши ключові терміни, формується часовий ряд, який в подальшому буде трансформований у ненаправлену мережу термінів. Далі здійснюється побудова ненаправленої мережі

термінів з використанням алгоритму побудови графа горизонтальної видимості на основі отриманого часового ряду. Наступним кроком є визначення напрямків зв’язків у отриманій на попередньому етапі ненаправленій мережі із термінів, що відповідають ключовим поняттям обраної предметної галузі – побудова так званого направленого графа горизонтальної видимості (Directed Horizontal Visibility Graph – DHVG) за одним із запропонованих правил визначення напрямків зв’язків. Після цього застосовується запропонований вище підхід до розрахунку вагових значень зв’язків у направленій мережі термінів – так званий направлений зважений граф горизонтальної видимості (Directed Weighted Horizontal Visibility Graph – DWHVG). Як результат, направлені зважені мережі термінів, побудовані за допомогою запропонованої технологічної схеми, можна використовувати як основу для автоматизованої побудови термінологічних онтологій предметних газузей з якими тематично пов’язані тексти.

Пропонується процес побудови БЗ СППР з використанням направлених зважених мереж термінів при розпізнаванні ІО. Припустимо, що є в наявності побудована з використанням контент моніторингу предметної галузі ІО достатньо якісна направлена зважена мережа термінів (достатнього обсягу, репрезентативна, без помилок, без надлишковості та з достатнім рівнем стійкості ваг). Це і буде початкова мережа термінів, яка буде використовуватися у подальшому. Далі пропонується наступна методика використання направлених зважених мереж термінів для формування БЗ СППР під час розпізнавання ІО:

1) Проводиться попередня побудова БЗ СППР з використанням контент моніторингу та експертної інформації, визначається ряд декомпозицій, в рамках яких потрібно визначити відповідні часткові коефіцієнти впливу (ЧКВ).

2) Виконується аналіз початкової направленої зваженої мережі термінів на предмет повноти покриття предметної галузі у відповідності до наявної попередньо побудованої БЗ СППР та вибраної декомпозиції.

3) Для кожної "покритої" декомпозиції визначається необхідний рівень абстракції та стратифікації і знаходяться відповідні терміни у початковій мережі, які відповідають кожному з об'єктів вибраної декомпозиції.

4) Формується нова мережа термінів шляхом об'єднання певних вузлів початкової мережі у відповідності до кожного з об'єктів вибраної декомпозиції (цілі та підцілей).

5) Знаходяться значення впливів підцілей на ціль шляхом об’єднання відповідних впливів. Отримані значення впливів нормуються та заносяться в БЗ СППР в якості відповідних ЧКВ.

6) Переходимо до пункту 3, доки не пройдемо по всіх "покритих" декомпозиціях.

Переваги запропонованого підходу полягають в наступному: економія часових та фінансових ресурсів за рахунок зменшення використання експертної інформації; можливість виявлення прогалин в БЗ СППР під час аналізу початкової направленої зваженої мережі термінів; об'єктивізація визначення ЧКВ.

Недоліками запропонованого підходу є: складність і, часом, неоднозначність знаходження відповідності деяких досить складних та широких цілей термінам мережі; відсутність можливості застосування підходу для інших сфер, окрім розпізнавання ІО.

Висновки до розділу 4

У цьому розділі були висвітлені результати практичного застосування запропонованої методики побудови мережевих моделей предметних галузей на основі текстових корпусів.

Спершу була представлена цілісна технологічна схема виокремлення та формування ключових термінів із текстів інформаційних повідомлень, яка передбачає послідовну обробку отриманого на вхід природомовного тексту за допомогою функцій NLP бібліотек мови програмування python, формування та виокремлення ключових слів, біграм та триграм за визначеними шаблонами, та їх подальше статистичне зважування за частотою появи у тексті.

У цьому розділі дисертаційної роботи був запропонований, реалізований та апробований лінгвостатистичний метод автоматичного екстрагування,

дослідження динаміки і виявлення взаємозв’язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів. Також запропонована форма візуального відображення інформаційного потоку в розрізі фразеологізмів і дат – Ph-Di діаграма (Phraseme Diagram).

У цьому розділі також були представлені модель середовища семантичного інформаційного пошуку та модель ранжування ранжування як окремих документів, так і джерел інформації, що стосуються визначеної у інформаційному запиті проблемної галузі. Також наведений приклад формування рейтингу джерел, щодо відповідності тематики інформаційного запиту на основі порівняння семантичних мереж.

Наприкінці розділу було розглянуто методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій.

ВИСНОВКИ

В результаті дисертаційного дослідження вирішено актуальне науково-практичне завдання, що стосується концептуалізації та подальшої формалізації у вигляді мережі термінів неструктурованих текстових даних, що містяться у тематичних інформаційних потоках.

На підставі проведеного аналізу сучасних мережевих підходів до структуризації текстових даних, що розподілені у динамічних інформаційних потоках мережі Інтернет, й огляду існуючих лінгвочтатистичних методів формування та аналізу мережевих моделей предметних галузей як мереж із текстів визначеної теми були виявлені основні проблеми формування й аналізу таких мереж. Це дозволило зробити висновок про актуальність й необхідність розробки нових та удосконалення існуючих лінгвостатистичних методів.

Тож у дисертаційній роботі отримано ряд результатів, що містять елементи наукової новизни, а саме:

1. запропоновано та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency) – глобальна частота терміна. Було встановлено, що на відміну від звичайного статистичного показника TF-IDF, GTF дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно-важливий термін зустрічається майже у кожному документі корпусу;

2. запропоновано метод виокремлення ключових термінів із текстового корпусу зі застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);

3. запропоновано лінгвостатистичний метод автоматичного екстрагування і виявлення взаємозв’язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів;

4. запропоновано форму візуального відображення інформаційного потоку в розрізі фразеологізмів – Ph-Di діаграму;

5. вперше запропоновано правила визначення напрямків зв’язків між вузлами ненаправленої мережі, сформованої зі ключових слів та словосполучень тематичного текстового масиву, що змістовно відноситься до певної предметної галузі;

6. запропоновано та розроблено новий метод визначення напрямків зв’язків між вузлами ненаправленої мережі, сформованої зі слів та словосполучень тематичного текстового масиву із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);

7. запропоновано та розроблено новий підхід до визначення вагових значень зв’язків у мережі термінів;

8. запропоновано цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусів;

9. запропоновано методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій;

10. запропоновано методику порівняння текстових документів та визначення ступеня їх подібності, що базується на формуванні та порівнянні відповідних їм семантичних мереж, та на основі цієї методики запропоновано модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Отримані результати мають значне практичне значення, оскільки мережеві моделі, сформовані на основі текстових корпусів, виявляються корисним інструментом для структуризації, візуалізації та аналізу текстових даних у різних предметних галузях. Ці моделі допомагають уточнити семантичну структуру термінів, полегшують розуміння зв'язків між поняттями та сприяють вивченню концептуальних зв'язків. Крім того, вони мають потенціал для покращення аналізу текстових даних, що є важливим для розвитку систем штучного інтелекту та систематизації інформації у різних сферах. Представлена мережева методика порівняння текстових документів може бути використана для виявлення структурних і термінологічних розбіжностей у правовій сфері, що сприятиме

парламентському контролю та гармонізації міжнародного права. Також такі моделі сприяють удосконаленню систем пошуку та рейтингування інформації, полегшуючи користувачам доступ до необхідної інформації та розуміння взаємозв'язків між поняттями у величезних обсягах даних. Крім того, вони відіграють важливу роль у створенні та управлінні базами знань у різних організаціях, сприяючи фільтрації та структуруванню великих обсягів інформації. Ці моделі можуть слугувати основою для прийняття ефективних рішень у різних галузях, від бізнесу до академічних досліджень, допомагаючи зрозуміти ключові аспекти та їх взаємозв'язки. Крім того, результати цих досліджень можуть сприяти подальшому розвитку технологій обробки природної мови та автоматичної обробки текстів, спрощуючи автоматизацію аналізу текстової інформації.

Отже, запропоновані та розроблені методи формування мережевих моделей предметних галузей на основі текстових корпусів мають широкий спектр застосувань і є ключовим інструментом для подальших досліджень та практичного використання у багатьох галузях.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ланде, Д. В., & Дмитренко, О. О. (2018). Створення мереж слів на основі текстів з використанням алгоритмів графів видимості. *Information Technology and Security: Ukrainian research papers collection, 2018, Vol. 6, Iss. 2 (11).* 5-18. DOI: doi.org/10.20535/2411-1031.2018.6.2.153486.
2. Lande, D. V., Dmytrenko, O. O., & Snarskii, A. A. (2018). Transformation texts into complex network with applying visability graphs algorithms. *Информационные технологии и безопасность*. *Матеріали XVIII Міжнародної науково-практичної конференції ІТБ-2018. - К.: ООО "Инжиниринг"*, 20-33.
3. Lande, D. V., Dmytrenko, O. O., & Snarskii, A. A. (2018). Transformation texts into complex network with applying visability graphs algorithms. *Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). In CEUR Workshop Proceedings (ceur-ws.org). Vol-2318,* (pp. 95-106) *urn:nbn:de:0074-2318-4*.
4. Ланде, Д. В., Дмитренко, О. О., & Радзієвська, О. Г. (2019). Побудова онтологій в галузі права за даними сервісу Google Scholar. *Інформація і право*, *1*(4), 74-85. DOI: doi.org/10.37750/2616-6798.2019.1(28).221313
5. Ланде, Д. В., & Дмитренко, О. О. (2019). Побудова мережі термів у сфері кібербезпеки за даними сервісу Google Scholar. *Матеріали XVII Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (25 - 26 квітня 2019 р., м. Київ, Україна)*, 143-145.
6. Дмитренко О.О. (2019). Створення термінологічних онтологій предметних областей на базі ресурсу Google Scholar. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» 16-17 травня 2019 року: збірник / - Київ: ІПРІ НАН України*, 108-109.
7. Ланде, Д. В., Дмитренко, О. О., & Радзієвська, О. Г. (2019). Визначення напрямків зв’язків у мережі термінів. *Інформаційні технології та безпека.*

Матеріали XIX Міжнародної науково-практичної конференції «ІТБ-2019». Київ: ТОВ «Инжиниринг», 103-112.

8. Lande, D., Dmytrenko, O., & Radziievska, O. (2019). Determining the directions of links in undirected networks of terms. *Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019). In CEUR Workshop Proceedings (ceur-ws.org). Vol-2577,* (pp. 132-145). ISSN 1613-0073.
9. Ланде, Д. В., & Дмитренко, О. О. (2019). Визначення вагових значень зв'язків у мережі термінів. *Реєстрація, зберігання і обробка даних*, *21(4),* 40-48. DOI: doi.org/10.35681/1560-9189.2019.21.4.199357
10. Lande, D. V., Dmytrenko, O. O., & Radziievska, O. H. (2020). Subject domain models of jurisprudence according to google scholar scientometrics data. *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference. Lviv, Ukraine, April 23-24, 2020. In CEUR Workshop Proceedings (ceur-ws.org).* - Vol-2604, (pp. 32-43). ISSN 1613-0073.
11. Ланде, Д. В., & Дмитренко, О. О. (2020). Метод побудови направлених зважених мереж термінів на основі текстових корпусів. *Матеріали XVIII Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики" (12 - 13 травня 2020, м. Київ, Україна)/ НТУУ "КПІ",* 68-71.
12. Ланде, Д. В., & Дмитренко, О. О. (2020). Побудова направлених зважених мереж термінів. *XI Всеукраїнська науково-практична конференція «Актуальні проблеми управління інформаційною безпекою держави»: зб. тез наук. доп. наук.-практ. конф. (Київ, 15 травня 2020 р.). [Електронне видання]. – Київ : НА СБУ*, 129-130.
13. Lande D. V., Dmytrenko O. O., Andriichuk O. V., Tsyganok V. V., & Porplenko Y. V. (2020). Building of directed weighted networks of terms for decision-making support during information operations recognition. *Математичне та імітаційне моделювання систем. МОДС 2020 : тези доповідей П'ятнадцятої міжнародної*

науково-практичної конференції (29 червня . 01 липня 2020 р., м. Чернігів) / М-во освіти і науки України ; Нац. Акад. наук України ; Академія технологічних наук України ; Інженерна академія України та ін. - Чернігів : ЧНТУ, 147-148.

14. Lande D. V., Dmytrenko O. O., Andriichuk O. V., Tsyganok V. V., & Porplenko Y. V. (2020). Building of directed weighted networks of terms for decision-making support during information operations recognition, *In: Mathematical Modeling and Simulation of Systems (MODS'2020). MODS 2020. Advances in Intelligent Systems and Computing, vol 1265*, (pp. 197-208). *Springer, Cham. Pages*. DOI: doi.org/10.1007/978-3-030-58124-4_19
15. Дмитренко О.О. (2020). Побудова мереж термінів на основі тематичних інформаційних публікацій. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція ІПРІ НАНУ «Реєстрація зберігання і обробка даних» 28-29 вересня 2020 року: збірник / - Київ: ІПРІ НАН України*, 107-108.
16. Lande D. V., & Dmytrenko O. O. (2020) Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora. *In 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC*) (pp. 1-4). *IEEE.* DOI: doi.org/10.1109/SAIC51296.2020.9239182
17. Ланде, Д. В., & Дмитренко, О. О. (2020). Методика виокремлення ключових слів і словосполучень та побудови направлених зважених мереж термінів із застосуванням Part-of-Speech tagging, Інформаційні технології і безпека. *Матеріали XX Міжнародної науково-практичної конференції ІТБ-2020. - Київ: Інжиніринг,* 140-144. ISBN: 978-966-2344-77-6
18. Lande, D., Andriichuk, O., Dmytrenko, O., Tsyganok, V., & Porplenko, Y. (2020). Побудова баз знань систем підтримки прийняття рішень з використанням направлених мереж термінів при дослідженні інформаційних операцій. *Information Technology and Security*, *8*(2), 153-163. DOI: doi.org/10.20535/2411-1031.2020.8.2.222597
19. Дмитренко, О. О. (2020). Побудова направлених зважених мереж термінів із застосуванням Part-of-speech tagging. *Реєстрація, зберігання і обробка даних, 22*(4), 47-55. DOI: doi.org/10.35681/1560-9189.2020.22.4.225914

20. Lande, D. V., & Dmytrenko, O. O. (2020). Methodology for Extracting of Key Words and Phrases and Building Directed Weighted Networks of Terms with Using Part-of-speech Tagging. *Selected Papers of the XX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2020) In CEUR Workshop Proceedings (ceur-ws.org). - Vol-2859* (pp. 168-177). ISSN 1613-0073
21. Ланде, Д. В., & Дмитренко, О. О. (2021). Побудова онтологічних моделей у галузі права. *Актуальні проблеми управління інформаційною безпекою держави: зб. тез наук. доп. наук.-практ. конф. (Київ, 26 березня 2021 р.). [Електронне видання]. - Київ : НА СБУ,* 62-63.
22. Ланде, Д. В., & Дмитренко, О. О. (2021). Формалізація знань та побудова термінологічних онтологій у правовій галузі. *Парламентський контроль в умовах децентралізації державної влади та цифрової трансформації в Україні: стан і проблеми: матеріали Першої всеукраїнської науково-практичної конференції, м. Київ, 30 березня 2021 р.*, 35-39.
23. Lande D.V., & Dmytrenko O.O. (2021). Using Part-of-Speech Tagging for Building Networks of Terms in Legal Sphere. *In Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference. Kharkiv, Ukraine, April 22-23, 2021. CEUR Workshop Proceedings (ceur-ws.org). - Vol-2870* (pp. 87-97). ISSN 1613-0073.
24. Ланде, Д. В., & Дмитренко, О. О. (2021). Використання направлених зважених мереж термінів для визначення ступеня подібності текстів. *Міжнародна наукова-технічна конференція "Інтелектуальні технології лінгвістичного аналізу": Тези доповідей. - К.: НАУ*, 7.
25. Zgurovsky, M. Z., Boldak, A. O., Lande, D. V., Yefremov, K. V., Pyshnograiev, I. O., Soboliev, A. M., & Dmytrenko, O. O. (2021). *Enhancing the Relevance of Information Retrieval in Internet Media and Social Networks in Scenario Planning Tasks, IEEE International Conference on System Analysis & Intelligent Computing SAIC 2021: System Analysis & Intelligent Computing. Studies in Computational Intelligence, vol 1022. Springer, Cham*, 187-199 DOI: https://doi.org/10.1007/978-3-030-94910-5_10

26. Ланде, Д. В., & Дмитренко, О. О. (2021). Побудова семантичних мереж та визначення ступеня розбіжності текстів. *Інформація і право*, *2*(41), 44-51. DOI: doi.org/10.37750/2616-6798.2022.2(41).270362
27. Dmytrenko, O. O., & Lande, D. V. (2022). Building of semantic networks to determine the degree of text similarity or difference. *Теоретичні і прикладні проблеми фізики, математики та інформатики: матеріали XX Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених (15 червня 2022 р., м. Київ, Україна). - Київ : КПІ ім. Ігоря Сікорського, Вид. "Політехніка"*, 197-202.
28. Lande, D., Soboliev, A., & Dmytrenko, O. (2022). Intelligent technologies in information retrieval systems. *Artificial intelligence*, 27(1), 260-268. DOI: https://doi.org/10.15407/jai2022.01.260
29. Lande, D. V., Dmytrenko, O. O., Shevchenko, A. I., Klymenko, M. S., & Vakulenko, M. O. (2023). Spoken language identification based on the transcript analysis. *Digital Scholarship in the Humanities*, 38(2), 586-595. DOI: https://doi.org/10.1093/llc/fqac052
30. Дмитренко О.О. (2022). Програмний модуль автоматичного екстрагування ключових термінів з інформаційних потоків. *Реєстрація, зберігання і обробка даних. Щорічна підсумкова наукова конференція 27-28 вересня 2022 року: збірник / - Київ: ІПРІ НАН України*, 122-123.
31. Zgurovsky, M. Z., Lande, D. V., Yefremov, K., Dmytrenko, O. O., Boldak, A. O., & Soboliev, A. M. (2022). Extracting and Identifying Relationships of Key Phrases in Information Flows. *In 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC) 04-07 October 2022*, (pp. 1-5). ISBN:979-8-3503-9674-4. DOI: 10.1109/SAIC57818.2022.9923019
32. Dmytrenko, O. (2022). Formation Networks of Terms for Identifying Semantic Similarity or Difference Degree of Texts in Cybersecurity. *Theoretical and Applied Cybersecurity*, 4(1). DOI: doi.org/10.20535/tacs.2664-29132022.1.274118
33. Zgurovsky, M. Z., Lande, D. V., Dmytrenko, O. O., Yefremov, K., Boldak, A. O., & Soboliev, A. M. (2022). Technological Principles of Using Media Content for

Evaluating Social Opinion. *System Analysis and Artificial Intelligence. Studies in Computational Intelligence*, Springer, Cham, 1107, 379-396 DOI: https://doi.org/10.1007/978-3-031-37450-0_22

34. Дмитренко, О. О. (2023). Формування та дослідження динамічних мереж термінів. Інформаційні технології і безпека. Матеріали XXIII Міжнародної науково-практичної конференції ІТБ-2023. - Київ: Інжиніринг, 83-84. ISBN: 978-966-2344-96-7.
35. Ланде Д. В., Дмитренко О. О., & Єфремов К. В. (2022). Комп'ютерна програма автоматичної побудови мереж термінів на основі аналізу текстових потоків «TermsNet». *Свідоцтво про реєстрацію авторського права на твір № с202204275 від 19.09.2022.* Державна організація «Український національний офіс інтелектуальної власності та інновацій».
36. Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
37. Большакова, Е. И., Клышинский, Э. С., Ландэ, Д. В., Носков, А. А., Пескова, О. В., & Ягунова, Е. В. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. ISBN 978.5.94506.294.8
38. Ланде, Д. В. (2017). Аналіз інформаційних потоків у глобальних комп’ютерних мережах (за матеріалами наукової доповіді на засіданні Президії НАН України 25 січня 2017 р.). *Вісник Національної академії наук України*, (3), 45-53. DOI: https://doi.org/10.15407/visn2017.03.045
39. Flux. (n.d.) *Humanity Doubles Its Data Creation Every 18 Months, And It Has Powerful Implications.* Retrieved November 22, 2020, from https://www.fluxmagazine.com/data-creation-powerful-implications/
40. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.
41. Netcraft (2021). *Web Server Survey*. https://news.netcraft.com/archives/category/web-server-survey/
42. Petrosyan, A. (2023, October 24). *Internet usage worldwide - Statistics & Facts Statista.* Statista. https://www.statista.com/topics/1145/internet-usage-worldwide/

43. Bianchi, T. (2022, November 15). *Most popular websites worldwide as of November 2022, by total visits.* Statista. https://www.statista.com/topics/1145/internet-usage-worldwide/
44. Dixon S. J. (2023, November 15). *Facebook: quarterly number of MAU (monthly active users) worldwide 2008-2023.* Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
45. ProHoster. (2020). *Обсяги інтернет-трафіку 10 березня досягли рекордних показників.* ProHoster. https://prohoster.info/uk/blog/novosti-interneta/obyomy-internet-trafika-10-marta-dostigli-rekordnyh-pokazatelej
46. Poinsignon, L. (2020, March 17). *On the shoulders of giants: recent changes in Internet traffic.* Cloudflare. https://blog.cloudflare.com/on-the-shoulders-of-giants-recent-changes-in-internet-traffic/
47. Gross, B. M. (1964). The managing of organizations: The administrative struggle.
48. Ландэ, Д. В., Снарский, А. А., & Безсуднов, И. В. (2009). *Интернетика*. URSS.
49. Власова, Г. В. (2006). Індексування як процес аналітико-синтетичної переробки інформації: навч. посіб. *навч. посіб./ГВ Власова*, 172.
50. Сукиасян, Э. Р. (2005). Координатное индексирование: выбор терминов индексирования и формирование поискового образа документа. *Библиотека*, (3), 42-44.
51. Кушнаренко, Н. М., & Удалова, В. К. (2006). Наукова обробка документів: підручник. *К.: Знання*.
52. Михайлов, А. М., Черный, А. И., & Гиляревский, Р. С. (1968). Основы информатики.
53. Manning, C. D. (2009). *An introduction to information retrieval*. Cambridge university press.
54. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463, No. 1999). New York: ACM press.
55. Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, *7*(3), 216-244.

56. Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.
57. Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic* (Vol. 4, pp. 1-12). New Jersey: Prentice hall.
58. Dubois, D. J. (1980). *Fuzzy sets and systems: theory and applications* (Vol. 144). Academic press.
59. Hüllermeier, E., & Chen, S. M. (2015). Fuzzy information retrieval: Recent approaches and future directions. *Fuzzy Sets and Systems*, 281, 83-101.
60. Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice* (Vol. 520, pp. 131-141). Reading: Addison-Wesley.
61. Salton, G. (1983). *Modern information retrieval.* McGraw-Hill.
62. Fuhr, N. (1992). Probabilistic models in information retrieval. *The computer journal*, *35*(3), 243-25.
63. Іванов, О. В. (2013). Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія : Соціологічні дослідження сучасного суспільства: методологія, теорія, методи, 1045*(30), 69-74.
64. Тарануха, В. Ю. (2014) *Інтелектуальна обробка текстів: навчальний посібник.* К.: електронна публікація на сайті факультету.
65. Дарчук, Н. П. (2008). *Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник.* К.: Видавничо-поліграфічний центр «Київський університет.
66. Балог, В. (2005). Сучасний стан української комп'ютерної лінгвістики. *Лексикографічний бюлетень*.
67. Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.
68. Reese, R. M., & Bhatia, A. (2018). *Natural Language Processing with Java: Techniques for building machine learning and neural network models for NLP*. Packt Publishing Ltd.
69. Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

70. Goldberg, Y. (2022). *Neural network methods for natural language processing*. Springer Nature.
71. Никоненко, А. О. (2011). Огляд комп'ютерно-лінгвістичних методів обробки природномовних текстів. *Штучний інтелект*. 3, 174-181.
72. Husain, M. S., & Beg, M. R. (2013). Word sense ambiguity: A survey. In *International Journal of Computer and Information Technology* (Vol. 2, No. 6, pp. 2279-0764). Department of IT, Integral University.
73. Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science, 8*, 44-50.
74. Бісікало, О. В., & Висоцька, В. А. (2016). Метод лінгвістичного аналізу україномовного комерційного контенту. *Вісник Національного університету Львівська політехніка. Серія: Інформаційні системи та мережі*, (854), 185-203.
75. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48).
76. Sparck Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *MCB University Press 60*, 493-502. DOI: 10.1108/eb026526
77. Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, *17*, 305-338. DOI: 10.1007/s00799-015-0156-0
78. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284. DOI:10.1080/01638539809545028
79. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.
80. Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2018). 2.6 Singular Value Decomposition. *Numerical Recipes in C: The Art of Scientific*

Computing, Cambridge, England: Cambridge University Press, *2*, 59-70. ISBN 0-521-43108-5.

81. Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *Europhysics Letters*, *57*(5), 759.
82. Дарчук, Н. (2013). Автоматичний синтаксичний аналіз текстів корпусу української мови. *Українське мовознавство*, (43), 11-19.
83. Гладун, А. Я., & Рогушина, Ю. В. (2016). Семантичні технології: принципи та практики. *К.: Універсаріум*.
84. Sowa, J. F. (1992). Semantic networks. *Encyclopedia of artificial intelligence*, *2*, 1493-1511.
85. Sowa, J. F. (Ed.). (2014). *Principles of semantic networks: Explorations in the representation of knowledge*. Morgan Kaufmann.
86. Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding* (pp. 35-82). Morgan Kaufmann.
87. Miller, G. A., & Fellbaum, C. (1991). Semantic networks of English. *Cognition*, *41*(1-3), 197-229.
88. Vakkari, P. (2000). Cognition and changes of search terms and tactics during task performance: A longitudinal case study. In *Content-Based Multimedia Information Access-Volume 1* (pp. 894-907).
89. Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.
90. Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, *18*(1), 22-31.
91. Liu, H., Gegov, A., & Cocea, M. (2015). *Rule based systems for big data: a machine learning approach* (Vol. 13). Heidelberg, Germany: Springer. DOI: 10.1007/978-3-319-23696-4
92. Ландэ, Д. В. (2005). Поиск знаний в Internet. Профессиональная работа. *М.: Издательский дом «Вильямс*, 105.
93. Roussopoulos, N. D. (1977). *A semantic network model of data bases*.

94. Faber, P., León Araúz, P., & Prieto Velasco, J. A. (2009). Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, *1*(1), 1-23.
95. Ландэ, Д. В., & Снарский, А. А. (2014). Подход к созданию терминологических онтологий. *Онтология проектирования*, 2 (12), 83-91.
96. Лукашевич, Н. В., Добров, Б. В., & Чуйко, Д. С. (2008, May). Отбор словосочетаний для словаря системы автоматической обработки текстов. In *Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции "Диалог* (pp. 339-344).
97. Филиппович, Ю. Н., & Прохоров, А. В. (2002). Семантика информационных технологий.
98. Cancho, R. F. I., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1482), 2261-2265. DOI: 10.1098/rspb.2001.1800
99. Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1485), 2603-2606. DOI: 10.1098/rspb.2001.1824
100. Caldeira, S. M., Petit Lobão, T. C., Andrade, R. F. S., Neme, A., & Miranda, J. V. (2006). The network of concepts in written texts. *The European Physical Journal B-Condensed Matter and Complex Systems*, *49*, 523-529. DOI: 10.1140/epjb/e2006-00091-3
101. Cancho, R. F. I.Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*(5). DOI: 10.1103/PhysRevE.69.051915
102. Ferrer i Cancho, R. (2005). The variation of Zipf's law in human language. *The European Physical Journal B-Condensed Matter and Complex Systems*, *44*, 249-257. DOI: 10.1140/epjb/e2005-00121-8
103. Motter, A. E., De Moura, A. P., Lai, Y. C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, *65*(6). DOI: 10.1103/PhysRevE.65.065102.

104. Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences*, *99*(3), 1742-1747.
105. Teodorescu, M. (2017). Machine Learning methods for strategy research. *Harvard Business School Research Paper Series*, (18-011).
106. Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science School of Engineering and Applied Science University of Pennsylvania Philadelphia, PA 19104.
107. Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313-330.
108. Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg. PA. USA* (pp. 152-155). DOI: doi:10.3115/974499.974526
109. Sacred-Texts. (n.d.). *The Hypertext Bible.* Retrieved April 27, 2020, https://sacred-texts.com/bib/index.htm
110. Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4* (pp. 320-332). Springer International Publishing.
111. NLTK. (n.d.). *NLTK Documentation.* Retrieved October 20, 2019, https://www.nltk.org/
112. Stanza. (n.d.). *Stanza – A Python NLP Package for Many Human Languages.* Retrieved July 16, 2021, https://stanfordnlp.github.io/stanza/index.html
113. Stanza. (n.d.). *Models.* Retrieved July 16, 2021, https://stanfordnlp.github.io/stanza/models.html#human-languages-supported-by-stanza
114. Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157-169.

115. Wartena, C., Brussee, R., & Slakhorst, W. (2010, August). Keyword extraction using word co-occurrence. In *2010 workshops on database and expert systems applications* (pp. 54-58). IEEE.
116. Luque, B., Lacasa, L., Ballesteros, F., & Luque, J. (2009). Horizontal visibility graphs: Exact results for random time series. *Physical Review E*, *80*(4). DOI: 10.1103/PhysRevE.80.046103
117. Gutin, G., Mansour, T., & Severini, S. (2011). A characterization of horizontal visibility graphs and combinatorics on words. *Physica A: Statistical Mechanics and its Applications*, *390*(12), 2421-2428. DOI: 10.1016/j.physa.2011.02.031
118. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V. (2013, November). 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* (pp. 209-215). IEEE. DOI: 10.1109/MICAI.2013.33
119. Lacasa, L., Luque, B., Ballesteros, F., Luque, J., & Nuno, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, *105*(13), 4972-4975. DOI: 10.1073/pnas.0709247105
120. Zou, Y., Donner, R. V., Marwan, N., Donges, J. F., & Kurths, J. (2019). Complex network approaches to nonlinear time series analysis. *Physics Reports*, *787*, 1-97.
121. Bondy, J. A. (1982). *Graph theory with applications*.
122. Godsil, C., & Royle, G. F. (2001). Algebraic graph theory. *Springer Science & Business Media*, *207*. DOI: 10.1007/978-1-4613-0163-9
123. Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604-632.
124. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117. DOI: 10.1016/S0169-7552(98)00110-X
125. Lande, D. (2014). Building of networks of natural hierarchies of terms based on analysis of texts corpora. *arXiv preprint arXiv:1405.6068*.
126. Ashliman, D. L. (n.d.). *Little Red Riding Hood.* Retrieved November 18, 2019, https://sites.pitt.edu/~dash/type0333.html

127. Gephi (n.d.). *The Open Graph Viz Platform.* Retrieved November 23, 2023, https://gephi.org/
128. Wikipedia (n.d.). *Біблія (Огієнко).* Retrieved May 15, 2020, https://uk.wikisource.org/wiki/Біблія_(Огієнко)
129. Böttcher, A., & Wenzel, D. (2008). The Frobenius norm and the commutator. *Linear algebra and its applications*, *429*(8-9), 1864-1885.
130. Toolan, M. J. (2013). *Narrative: A critical linguistic introduction*. Routledge.
131. Tekin, F., & Meissner, V. (2022). Political Differentiation as the End of Political Unity? A Narrative Analysis. *The International Spectator*, *57*(1), 72-89. DOI: 10.1080/03932729.2022.2018823
132. Naiemi, F., Ghods, V., & Khalesi, H. (2022). Scene text detection and recognition: a survey. *Multimedia Tools and Applications*, *81*(14), 20255-20290. DOI: 10.1007/s11042-022-12693-7
133. Zhong, H., Ning, Z., Li, G., & Li, Z. (2022). A method of core concept extraction based on semantic-weight ranking. *Concurrency and Computation: Practice and Experience*, *34*(1)/ DOI: 10.1002/cpe.6504
134. Giatrakos, N., Alevizos, E., Artikis, A., Deligiannakis, A., & Garofalakis, M. (2020). Complex event recognition in the big data era: a survey. *The VLDB Journal*, *29*, 313-352. DOI: 10.1007/s00778-019-00557-w
135. Zgurovsky, M., Boldak, A., Lande, D., Yefremov, K., Pyshnograiev, I., Soboliev, A., & Dmytrenko, O. (2020, October). Enhancing the Relevance of Information Retrieval in Internet Media and Social Networks in Scenario Planning Tasks. In *IEEE International Conference on System Analysis & Intelligent Computing* (pp. 187-199). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-94910-5_10
136. Lande, D., & Yagunova, E. (2012). Dynamic frequency features as the basis for the structural description of diverse linguistic objects. In *CEUR Workshop Proceedings. Vol-934* (pp. 150-159).
137. Lande, D., Subach, I., Puchkov, O., & Soboliev, A. (2021). A Clustering Method for Information Summarization and Modelling a Subject Domain. *Information & Security*, *50*(1), 79-86. DOI: 10.11610/isij.5013

ДОДАТОК 1. СВІДОЦТВО ПРО РЕЄСТРАЦІЮ АВТОРСБКОГО ПРАВА НА ТВІР

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 115924

Комп'ютерна програма автоматичної побудови мереж термінів на основі аналізу текстових потоків «TermsNet»

(вид, назва твору)

Автор(и) **Ланде Дмитро Володимирович, Дмитренко Олег Олександрович, Єфремов Костянтин Вікторович**

(повне ім'я, псевдонім (за наявності))

Дата реєстрації 19 січня 2023 р.

Директор
Державної організації
«Український національний
офіс інтелектуальної власності
та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

ДОДАТОК 2. ДОВІДКА ПРО ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

УКРАЇНА

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

НАВЧАЛЬНО–НАУКОВИЙ ЦЕНТР
«СВІТОВИЙ ЦЕНТР ДАНИХ З ГЕОІНФОРМАТИКИ ТА СТАЛОГО РОЗВИТКУ»

03056, м. Київ, пр. Берестейський, 37; тел. (+38 044) 204 8014 тел./факс (+38 044) 204 8153
web: http://wdc.org.ua, e-mail: mail@wdc.org.ua

15.02.2024 № 0211/24
на № _____ від _____

АКТ

про впровадження результатів дисертаційної роботи
Дмитренка Олега Олександровича
на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу», поданої на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки»

Навчально-науковий центр «Світовий центр даних з геоінформатики та сталого розвитку» КПІ ім. Ігоря Сікорського (СЦД-Україна) веде багаторічну дослідну роботу за напрямком сценарного моделювання кризових і безпекових явищ та вивчення їх впливу на економіку і суспільство. Ці дослідження проводяться з використанням великих масивів неструктурованих даних різної природи з різних інформаційних джерел. Роботу Олега Дмитренка спрямовано на вирішення складних проблем, пов'язаних з обробкою та аналізом великих обсягів неструктурованих текстових даних в середовищі електронних медіа-ресурсів. В умовах сучасної війни великі обсяги цифрового контенту та висока швидкість поширення інформації створюють небезпечні виклики для суспільства, зокрема цілеспрямованого інформаційного впливу за допомогою електронних медіа-ресурсів на суспільну свідомість, настрої і поведінку людей, що вимагає наукового підходу до розробки ефективних стратегій протидії, а також нових методів та інструментів для виявлення таких інформаційних впливів. Це визначило нагальність й актуальність теми дослідження та потребу в подальших науково-прикладних розробках.

В дисертаційній роботі розроблено нові методи формування мережевих моделей предметних галузей на основі текстових корпусів та розроблено нові методи аналізу побудованих мереж з метою прийняття ефективних рішень у відповідних предметних галузях, з якими змістовно пов'язані тексти.

Результати дисертаційної роботи Дмитренка Олега Олександровича на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу» було використано і впроваджено при реалізації програмно-технічних засобів в середовищі Інформаційно-аналітичного ситуаційного центру КПІ ім. Ігоря Сікорського для комплексного використання та інтелектуального аналізу великих масивів неструктурованих даних різної природи, в т.ч. результатів обробки текстів природної мови, в ході виконання низки держбюджетних і договірних науково-дослідних робіт (НДР) та проектів ННЦ «СЦД-Україна», серед яких:

- НДР «Створення інтегрованої платформи для ситуаційного аналізу соціально-економічних і безпекових явищ» (0121U113470), в рамках якої було впроваджено нові методи лінгвостатистичного аналізу надвеликих масивів текстових даних для оцінювання ставлення суспільства до дій влади на основі аналізу даних з відкритих Інтернет-видань і соціальних мереж.
- НДР «Створення інформаційно-аналітичного ситуаційного центру для сценарного моделювання кризових і безпекових явищ та вивчення їх впливу на економіку і суспільство» (0121U109764), в рамках якої було впроваджено метод формування мережевих моделей предметних галузей (семантичних мереж) на основі текстових корпусів.
- НДР «Розробка методології та програмно-технічного комплексу для системної оцінки безпекового рівня територій України на основі супутникових даних за умов множинних військових загроз» (0123U102015), в рамках якої впроваджено метод виокремлення ключових термінів із текстів із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging).
- НДР «Розробка програмно-технічного комплексу інтелектуального аналізу неструктурованих даних методами штучного інтелекту та OSINT для планування військових операцій» (0124U000838), в рамках якої впроваджено методику порівняння текстових документів (новинних повідомлень в середовищі електронних медіа-ресурсів), що базується на побудові та порівнянні відповідних їм семантичних мереж, та на основі цієї методики впроваджено модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Окрім того, впровадження нових методів лінгвостатистичного аналізу надвеликих масивів текстових даних та нового методу виокремлення ключових термінів із текстових (новинних) повідомлень для формування їх семантичних мереж й вимірювання семантичної близькості відповідних повідомлень в середовищі електронних медіа-ресурсів до інформаційної системи аналітичної обробки інформації Інформаційно-аналітичного ситуаційного центру КПІ ім. Ігоря Сікорського дозволило покращити інтелектуальну обробку та аналіз текстових даних з веб-сайтів та соціальних мереж під час формування дайджестів та отримання нових інсайтів, а реалізація моделі ранжування як окремих документів, так і джерел інформації у інформаційній системі збору та аналітичного оброблення дозволила вцілому покращити рейтингування джерел на 12%. Також вищезгадані впровадження допомогли підвищити повноту охоплення інформації на 25% за рахунок врахування ширшого переліку інтернет-джерел та медіа-ресурсів, зокрема таких, як канали Telegram і Youtube, а також підвищити загальну швидкість обробки текстових даних та оперативність надання релевантної інформації замовникам у відповідь на їх запит під час інформаційного пошуку.

Директор
к.т.н., доц. Костянтин ЄФРЕМОВ