УДК 004.7:001.8

ЛАНДЕ Д.В., *доктор технічних наук,* керівник наукового центру
НДІІП НАПрН України
ДМИТРЕНКО О.О., аспірант, Інститут проблем реєстрації інформації НАН України
РАДЗІЄВСЬКА О.Г., *кандидат юридичних наук,* старший науковий співробітник
НДІІП НАПрН України

ПОБУДОВА ОНТОЛОГІЙ В ГАЛУЗІ ПРАВА ЗА ДАНИМИ СЕРВІСУ GOOGLE SCHOLAR

Анотація. *У статті викладені підходи до структуризації даних, розподілених в наукових документальних ресурсах мережі Інтернет. Представлено методи формування моделей предметних галузей як мереж із термінів певної тематики, які є інформаційно важливими в межах заданої теми. Побудовано мережі природніх ієрархій термінів для корпусу документів, пов'язаних з тематиками "Criminal Law" та "Copyright Law". Розглянута у статті методика створення мережі зі слів та словосполучень – алгоритм формування мереж природніх ієрархій термінів сприятиме формуванню й удосконаленню понятійного і термінологічного апарату у правовій сфері та гармонізації національного і міжнародного права.*

Ключові слова: *інформаційні ресурси, правова інформація, термінологія, мережа природної ієрархії термінів, предметна область, онтологія.*

Summary. *The article presents approaches to the structuring of data distributed in scientific documentary resources of the Internet. It represents generating methods of subject branches models as networks in terms of certain subjects, containing important information in the framework of the given topic. Networks of natural hierarchies of terms for the corpus of documents related to the topics "Criminal Law" and "Copyright Law" are built. The considered methodic of creating a network of words and phrases, the implementation of the algorithm for the formation of networks of natural hierarchies of terms – will contribute to the formation and improvement of conceptual and terminological apparatus in the legal sphere and the harmonization of national and international law.*

Keywords: *information resources, legal information, terminology, network of natural hierarchy of terms, subject domain, ontology.*

Аннотация. *В статье изложены подходы к структуризации данных, распределенных в научных документальных ресурсах сети Интернет. Представлены методы формирования моделей предметных областей как сетей из терминов определенной тематики, информационно важных в пределах заданной темы. Построены сети естественных иерархий терминов для корпуса текстовых документов связанных с тематиками "Criminal Law" и "Copyright Law". Рассмотренная в статье методика создания направленной сети со слов и словосочетаний – алгоритм формирования сетей естественных иерархий терминов способствует формированию и совершенствованию понятийного и терминологического аппарата в правовой сфере и гармонизации национального и международного права.*

Ключевые слова: *информационные ресурсы, правовая информация, терминология, сеть естественной иерархии терминов, предметная область, онтология.*

Постановка проблеми. Глобалізація інформаційного простору та стрімкий розвиток інформаційно-комунікаційних технологій призвели до не менш стрімкого розвитку інформаційних ресурсів. Виникла невідкладна потреба у дослідженнях та розробках нових методів та засобів більш швидкого пошуку та синтезу потрібної інформації, нових підходів до створення ефективних пошукових систем. Також постає питання зручного

візуального представлення отриманої інформації. Сучасний науково-технічний прогрес породжує нові суспільні відносини, а також суттєво трансформує існуючі. Це значно ускладнює процеси своєчасного виявлення найбільш важливих суспільних відносин та встановлення правовідносин. Структуризація даних, розподілених в інформаційних ресурсах, методами формування мереж із текстів певної тематики на основі автоматично екстрагованих ключових термінів, допоможе спростити поставлене завдання та сприятиме формуванню й удосконаленню понятійного і термінологічного апарату у правовій сфері та гармонізації національного і міжнародного права.

Дуже важливим етапом у комплексних дослідженнях є детальне формалізоване представлення знань обраної предметної області (Subject Domain), придатне для автоматизованої обробки – створення онтологій, в тому числі й правових. Процес побудови великих тематичних онтологій зазвичай є складним та ресурсозатратним. Окремий крок такої формалізації – це 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 [1; 2]. Досліджуючи лексику, яка використовується в певних текстових масивах, за окремими ключовими термінами-маркерами можна визначати відповідність цих текстів до певної тематики загальних інформаційних потоків. Не менш складною й відкритою проблемою концептуалізації є встановлення зв'язків між термінами.

Також виникає питання щодо подальшого візуального представлення предметних областей. Однією із моделей предметних областей може розглядатися мережа слів (Language Network), вузли якої відповідають окремим поняттям, а ребра – зв'язкам між ними [3]. У цій статті описані підходи до формування мережевих структур із корпусу текстових документів на основі вибраних ключових термінів, які є інформаційно важливими в межах обраної теми.

Одним із методів створення термінологічних онтологій є алгоритм формування направленої мережі зі слів та словосполучень – алгоритм формування мереж природніх ієрархій термінів [4] для корпусу текстових документів. Цей алгоритм базується на використанні інформаційно-важливих елементів тексту, опорних слів та словосполучень (уніграм, біграм та триграм) [5], методика виявлення яких представлена в роботі [4]. Алгоритм створення мереж природніх ієрархій термінів передбачає побудову компактифікованого графу горизонтальної видимості [6 – 10] для термів – окремих слів, біграм та триграм, та встановленні направлених зв'язків між термами.

Як зазначено у роботі [11], алгоритм формування мереж природніх ієрархій термінів можна представити у вигляді послідовних етапів, які охоплюють попередню обробку отриманого корпусу текстових документів, виділення ключових слів та словосполучень, що є інформаційно-важливими в межах розглянутої предметної області, побудова компактифікованого графу горизонтальної видимості (CHVG), перерахунок сортування вагових значень виділених термів за обраним ваговим критерієм та вибір із них найбільш вагомих. Кінцевим етапом є безпосереднє формування мережі природніх ієрархій термінів (з'єднання вузлів зв'язками "входження") та її відображення.

Метою статті є побудова мережі природніх ієрархій термінів для корпусу текстових документів тематично пов'язаних з "Criminal Law" та "Copyright Law".

Виклад основного матеріалу

Формування корпусу текстових документів. Початковим етапом формування мережі термів, пов'язаної з певною предметною областю, є формування корпусу текстових документів. Для проведення досліджень була використана вільна доступна пошукова

система, яка індексує повний текст наукових публікацій – Google Scholar (https://scholar.google.com). На цьому етапі було вивантажено анотації перших 385-ти статей за запитом “Criminal Law” та анотації перших 490-та статей за запитом “Copyright Law”.

Обробка текстових документів та виокремлення ключових термів. На цьому етапі проводиться процес попереднього лексичного аналізу – розбиття тексту на елементарні одиниці (токени або леми). Токенізація (лематизація) є зазвичай початковим етапом обробки текстів, адже дає змогу працювати зі словом як з окремою сутністю, при цьому знаючи його контекст [12].

Наступним крок – зважування термів. В якості вагових значень термів, для формування часового ряду в якості функції, яка ставить у відповідність слову число, в даному дослідженні використовується класичний статистичний ваговий показник TF–IDF (з англ. Term Frequency – “частота слова”, Inverse Document Frequency – “обернена частота документа”) [13], хоча це не єдиний можливий для вирішення завдання виділення ключових термів підхід [4]. Цей статистичний ваговий показник використовується для оцінки важливості слів у контексті документа, що є частиною колекції документів чи корпусу [14]. Вага (значимість) слова пропорційна кількості вживань цього слова у документі і обернено пропорційна частоті вживання слова у інших документах колекції. Показник TF–IDF використовується в задачах аналізу текстів та інформаційного пошуку. Його можна застосовувати як один з критеріїв релевантності документа до пошукового запиту [15].

TF – відношення числа входжень обраного слова до кількості слів у документі. Таким чином, оцінюється важливість слова t_i в межах обраного документа. Термін введений Карен Спарк Джонс [16].

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

де: n_i – число входжень слова в документ;

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

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

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

де: $|D|$ – кількість документів колекції;

$|(d_i \supset t_i)|$ – 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:

$$TF–IDF = TF \circ IDF.$$

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

Беручи до уваги той факт, що в даному дослідженні розглядаються документи, що описують одну предметну область, то для запобігання втрати інформаційно-важливих елементів тексту, опорних слів та словосполучень (біграм та триграм), в якості

статистичного показника важливості терма було використано лише показник TF. Такий вибір пояснюється тим, що терми, які є ключовими для розглянутої предметної області й зустрічаються у більшості документів, матимуть низьке числове значення IDF (отже, і низьким буде числове значення TF–IDF), в той час, коли насправді ці слова є інформаційно-важливими, тобто такими, що визначають структуру тексту.

Також щоб уникнути ситуації, що виникає під час роботи з текстовим корпусом заздалегідь визначеної тематики, коли інформаційно-важливий терм зустрічається майже у кожному документі корпусу і має низький ваговий показник TF, було застосовано глобальний TF–GTF (Global TF) [17].

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

де: n_i – загальна кількість появи терма i у всіх документах корпусу;

$\sum_k n_k$ – загальна кількість термів у документах корпусу.

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

Вилучення стоп-слів. Також після етапу обробки текстових документів та виокремлення ключових термів в даному дослідженні пропонується вилучити стоп-слова, які не мають ніякого смислового навантаження, тобто є інформаційно-неважливими, а також біграми, які містять принаймні одне стоп-слово, та триграми, які починаються, або закінчуються стоп-словом. Стоп-словник, який використовувався в межах даного дослідження, був сформований на основі різних стоп-словників, які доступні за посиланнями:

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

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

Процес стематизації. Для об'єднання (злиття) слів, які мають спільний корінь в даному дослідженні здійснюється процес стематизації – скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс [18]. Стематизація або стемінг (англ. – stemming) є процесом нормалізації тексту шляхом знаходження основи слова. Варто зазначити, що основа слова не обов'язково співпадає з морфологічним коренем слова.

Існує декілька типів алгоритмів стемінгу, які розрізнюються відносно продуктивності, точності та відносно того, як долаються проблеми стемінгу [19].

Алгоритм стемінгу Мартіна Портера [20; 21] набув значного поширення та став де-факто стандартним алгоритмом стемінгу для англійської мови. Для проведення досліджень було використано стример, що реалізований на мові Python (бібліотека NLTK – Natural Language Toolkit).

Загалом, стемінг застосовується в лінгвістичній морфології та в пошукових системах для розширення пошукового запиту користувача та є частиною нормалізації тесту.

Алгоритм побудови компактифікованого графу видимості. У роботах [4; 6 – 8, 22; 23] запропоновано алгоритм побудови мереж термів – алгоритм побудови компактифікованого графу горизонтальної видимості (Compactified Horizontal Visibility Graph – CHVG). Загалом, мережа термів з використанням алгоритму горизонтальної видимості будується у три етапи. На першому етапі на горизонтальній осі позначається ряд вузлів, кожен з яких відповідає словам у тому порядку, в якому вони з'являються в тексті, а по вертикальній осі відкладаються вагові значення – числові оцінки. На

другому етапі будується граф горизонтальної видимості [9; 10]. Більш формально ідею побудови графу горизонтальної видимості можна представити наступним чином: два вузли t_i і t_j, які відповідають елементам часового ряду x_i і x_j, знаходяться у горизонтальній видимості тоді й тільки тоді, коли

$$x_k < \min\{x_i, x_j\}$$
$$\text{для всіх } t_k \ (t_i < t_k < t_j).$$

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

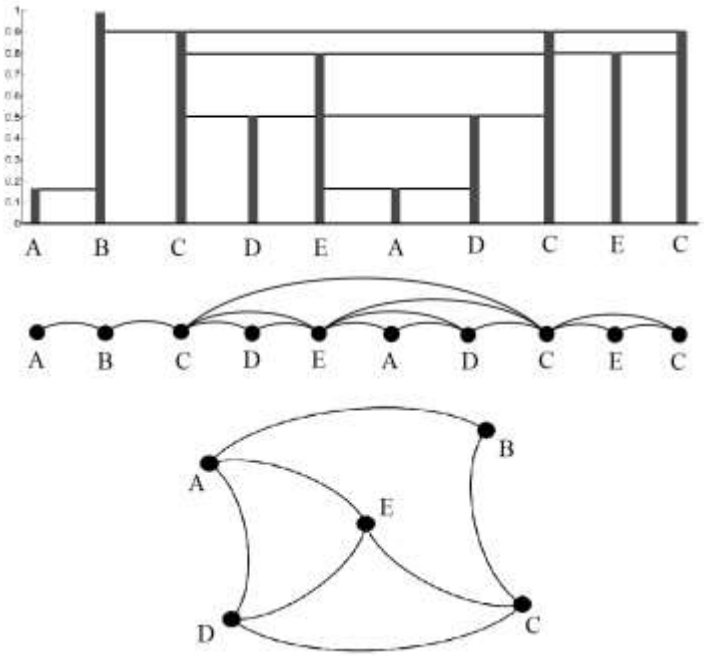

Рис. 1. Етапи побудови компактифікованого графу горизонтальної видимості

Формування мережі природніх ієрархій термінів. Для послідовностей термів (слів, біграм та триграм) та їх вагових значень, визначених за допомогою статистичного показника важливості терма – GTF, будуються компактифіковані графи горизонтальної видимості (CHVG).

Наступним кроком є перерахунок вагових значень, що відповідають термам у CHVG. Ця процедура дозволяє врахувати в подальшому також ті терми, які мають велике значення для загальної тематики текстового корпусу [22]. Під час виконання досліджень перерахунок ваг здійснюється з використанням алгоритму HITS [24; 25], завдяки якому визначається авторство чи посередництво для кожного вузла CHVG. Вибір форми вагового значення (авторство чи посередництво) немає значення, оскільки граф є ненаправленим Після цього всі терми упорядковуються за спаданням розрахованих вагових значень відповідних їм вузлів у CHVG.

Далі експертним методом визначається необхідний розмір (число N) створюваної мережі природніх ієрархій термінів, після чого вибирається N простих слів, біграм та триграм (всього $N+N+N$ елементів), що мають найбільші значення вагових показників відповідних їм вузлів у CHVG.

На наступному етапі будується сама мережа природніх ієрархій термінів, в якій вузли відповідають відібраним термам, а зв'язки між ними – входженням одного терма в інший (Рис. 2).

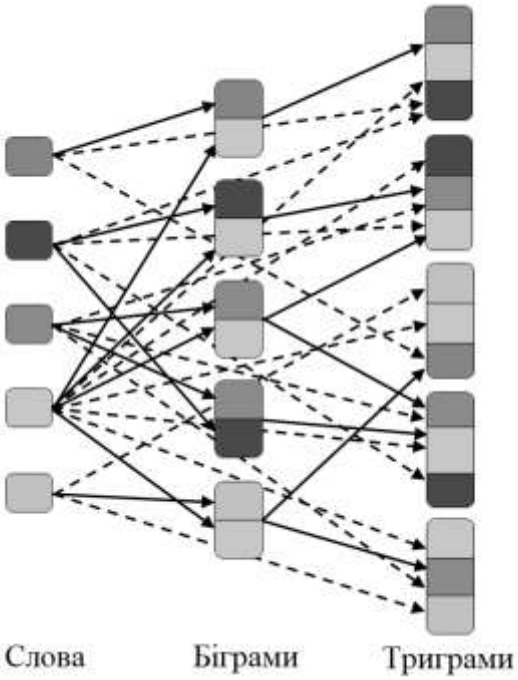

Рис. 2. Трирівнева модель мережі природніх ієрархій термінів

Заключним є відображення створеної мережі природніх ієрархій термінів засобами візуалізації графів. На вхід таким засобам подається матриця інцидентності у форматі csv, створена на етапі формування мережі природніх ієрархій термінів.

Мережа природніх ієрархій термінів, що створюється повністю автоматично, може розглядатися як основа для подальшого автоматизованого формування термінологічних онтологій за участю експертів.

Візуалізація й аналіз результатів дослідження. Для проведення досліджень в даній роботі використано корпус заздалегідь вибраних текстових документів, що тематично пов'язані з актуальною предметною областю – "Criminal Law". Імпортувавши стример, що реалізований на мові Python (бібліотека NLTK – Natural Language Toolkit), було попередньо здійснено процес стематизації текстового корпусу отриманого для запиту "Criminal Law" об'ємом 385 документи, внаслідок чого слова, які мають спільний корінь, були об'єднанні.

У Табл. 1 наведені списки найбільш вагомих термів (слів, біграм та триграм) для досліджуваної предметної області відповідно до мережевого рангового критерію HITS [24; 25].

Таблиця 1. Списки найбільш вагомих термів (слів, біграм та триграм) для "Criminal Law"

№	Слова	Біграми	Триграми
1	studi	restor_justic	civil_and_crimin
2	moral	onlin_librari	heinonlin_thi_articl
3	work	law_reform	crimin_law_doctrin
4	law	crimin_sanction	law_and_crimin

5	principl	columbia_law	theori_of_crimin
6	respons	crimin_respons	crimin_law_text
7	subject	civil_law	univers_of_pennsylvania
8	liabil	univers_press	american_crimin_justic
9	human	corpor_crimin	crime_and_crimin
10	right	gener_principl	case_and_materi
11	intern	articl_examin	analysi_of_crimin
12	role	case_involv	intern_and_compar
13	concept	mental_disord	principl_of_crimin
14	gener	compar_crimin	substant_crimin_law
15	univers	court_icc	american_crimin_law
16	examin	crimin_justic	crimin_court_icc
17	practic	feder_crimin	crime_against_human
18	polici	substant_crimin	field_of_crimin
19	year	intern_crimin	sanctiti_of_life
20	theori	crimin_code	philosophi_of_crimin
21	crime	war_crime	columbia_law_review
22	review	common_market	crimin_law_theori
23	case	compar_law	law_sj_schulhof
24	develop	paper_examin	law_and_criminolog
25	legal	sexual_violenc	journal_of_intern
26	heinonlin	law_review	crime_and_punish
27	articl	intern_law	english_crimin_law
28	public	social_scienc	role_of_crimin
29	social	intern_crime	law_and_procedur
30	histori	law_theori	crimin_law_case
31	court	american_law	intern_crimin_justic
32	author	soviet_crimin	wiley_onlin_librari
33	procedur	crimin_liabil	crimin_law_defens
34	american	crimin_procedur	intern_crimin_court
35	major	crimin_tribun	crimin_law_volum
36	discuss	crimin_law	crimin_law_enforc
37	question	feder_court	harvard_law_review
38	prosecut	intern_commun	pennsylvania_law_review
39	crimin	law_journal	crimin_and_civil
40	issu	secur_council	crim_l_criminolog
41	edit	law_enforc	compar_crimin_law
42	societi	law_doctrin	intern_crimin_law
43	defens	american_crimin	crimin_law_review
44	enforc	american_journal	journal_of_law
45	punish	human_right	intern_crimin_tribun
46	doctrin	militari_tribun	yale_law_journal
47	feder	mental_disabl	feder_crimin_law
48	justic	crimin_court	crimin_law_reform
49	problem	common_law	corpor_crimin_liabil
50	rule	prosecutori_discret	heinonlin_crimin_law

Використовуючи засоби програмного забезпечення для моделювання та візуалізації графів – Gephi (https://gephi.org) [26] побудована мережа природніх ієрархій термінів розміром 50+50+50 була візуалізована (Рис. 3).

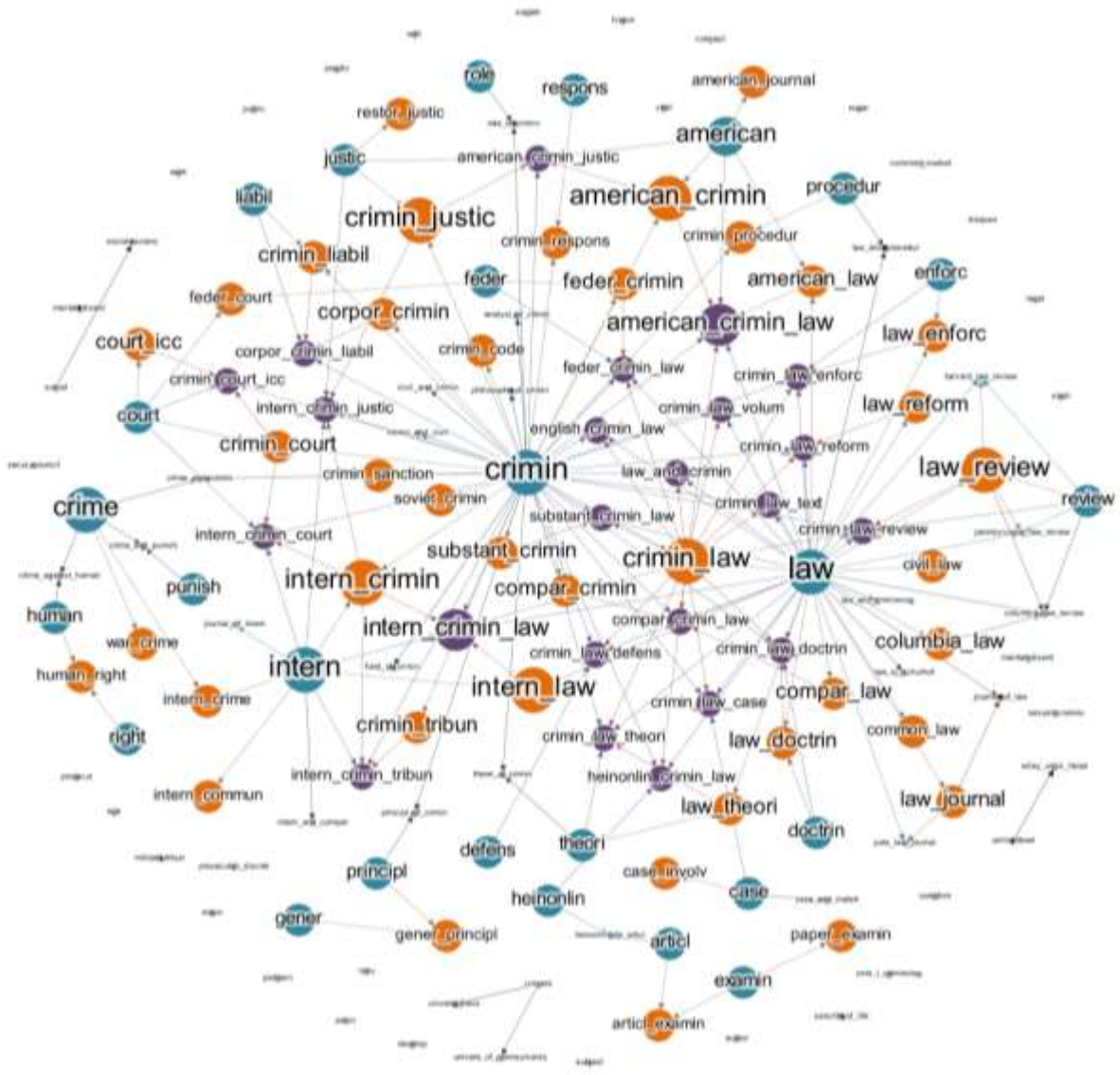

Рис. 3. Мережа природніх ієрархій термінів розміром 50+50+50
для предметної області “Criminal Law”

Також за допомогою засобів програмного забезпечення Gephi були отримані такі параметри створеної мережі: кількість вузлів – 150; кількість зв’язків – 205; щільність мережі – 0.009; кількість зв’язаних компонент – 35; середня довжина шляху – 1; середній коефіцієнт кластеризації – 0.121.

За топологічною особливістю побудована мережа має малий середній коефіцієнт кластеризації. Це пояснюється наявністю в мережі великої кількісті понять, сусіди яких мало пов’язані один з одним – це є ознакою так званих квазіієрархічних мереж. Невелика середня довжина шляху свідчить про те, що ця мережа також є “малим світом” (Small World) [27].

Також було досліджено корпус текстових документів, що тематично пов’язані з актуальною предметною областю – “Copyright Law”. Попередньо здійснено процес стематизації корпусу, що був отриманий для запиту “Copyright Law”, обсягом 490 документи.

В Табл. 2 наведені списки найбільш вагомих термів (слів, біграм та триграм) для досліджуваної предметної області відповідно до мережевого рангового критерію HITS.

Таблиця 2. Списки найбільш вагомих термів (слів, біграм та триграм) для “Copyright Law”

№	Слова	Біграми	Триграми
1	digit	copyright_legisl	type_of_tumour
2	intellectu	digit_technolog	wiley_onlin_librari
3	unit	properti_right	author_and_publish
4	music	copyright_protect	heinonlin_thi_articl
5	work	cardozo_art	fair_use_doctrin
6	law	law_reform	german_copyright_law
7	origin	violat_fall	law_in_canada
8	public	canadian_copyright	soc_y_usa
9	fair	three-step_test	access_to_copyright
10	protect	unauthor_copi	case_and_materi
11	properti	legal_studi	law_of_copyright
12	patent	intel_prop	intellectu_properti_right
13	right	intellectu_properti	purpos_of_copyright
14	intern	copyright_handbook	law_and_econom
15	creativ	univers_press	professor_of_law
16	analysi	berkeley_tech	literari_and_artist
17	current	digit_media	digit_millennium_copyright
18	gener	paid_violat	law_jc_ginsburg
19	inform	copyright_law	patent_and_copyright
20	nation	copyright_work	canadian_copyright_law
21	internet	septemb_9	law_is_base
22	theori	copyright_owner	law_a_commentari
23	question	subject_matter	aspect_of_copyright
24	copi	public_domain	public_or_part
25	artist	intern_copyright	librarian_and_educ
26	case	part_thereofispermit	law_of_septemb
27	develop	current_copyright	notion_of_origin
28	author	special_case	european_copyright_law
29	media	copyright_limit	transit_ma_schlosshauer
30	econom	law_review	republ_of_china
31	legal	digit_millennium	analysi_of_copyright
32	industri	american_copyright	intellectu_properti_law
33	creat	exclus_right	paid_violat_fall
34	articl	german_copyright	nation_inform_infrastructur
35	social	copyright_infring	protect_by_copyright
36	court	copyright_case	law_and_practic
37	publish	european_copyright	heinonlin_copyright_law
38	american	unfair_competit	approach_to_copyright
39	number	wto_panel	law_c_geiger
40	art	digit_copyright	version_and_permiss
41	materi	open_sourc	american_copyright_law
42	univers	law_journal	digit_copyright_law
43	copyright	properti_law	current_copyright_law
44	technolog	moral_right	springer-verlag_berlin_heidelberg
45	societi	copyright_fee	berlin_heidelberg_gmbh
46	softwar	copyright_soc	law_a_propos

47	cultur	copyright_issu	law_p_samuelson
48	heinonlin	price_discrimin	intern_copyright_law
49	limit	current_version	dichotomi_in_copyright
50	futur	law_school	guid_to_copyright

Використовуючи засоби програмного забезпечення для моделювання та візуалізації графів – Gephi побудована мережа природніх ієрархій термінів розміром 50+50+50 була візуалізована (Рис. 4).

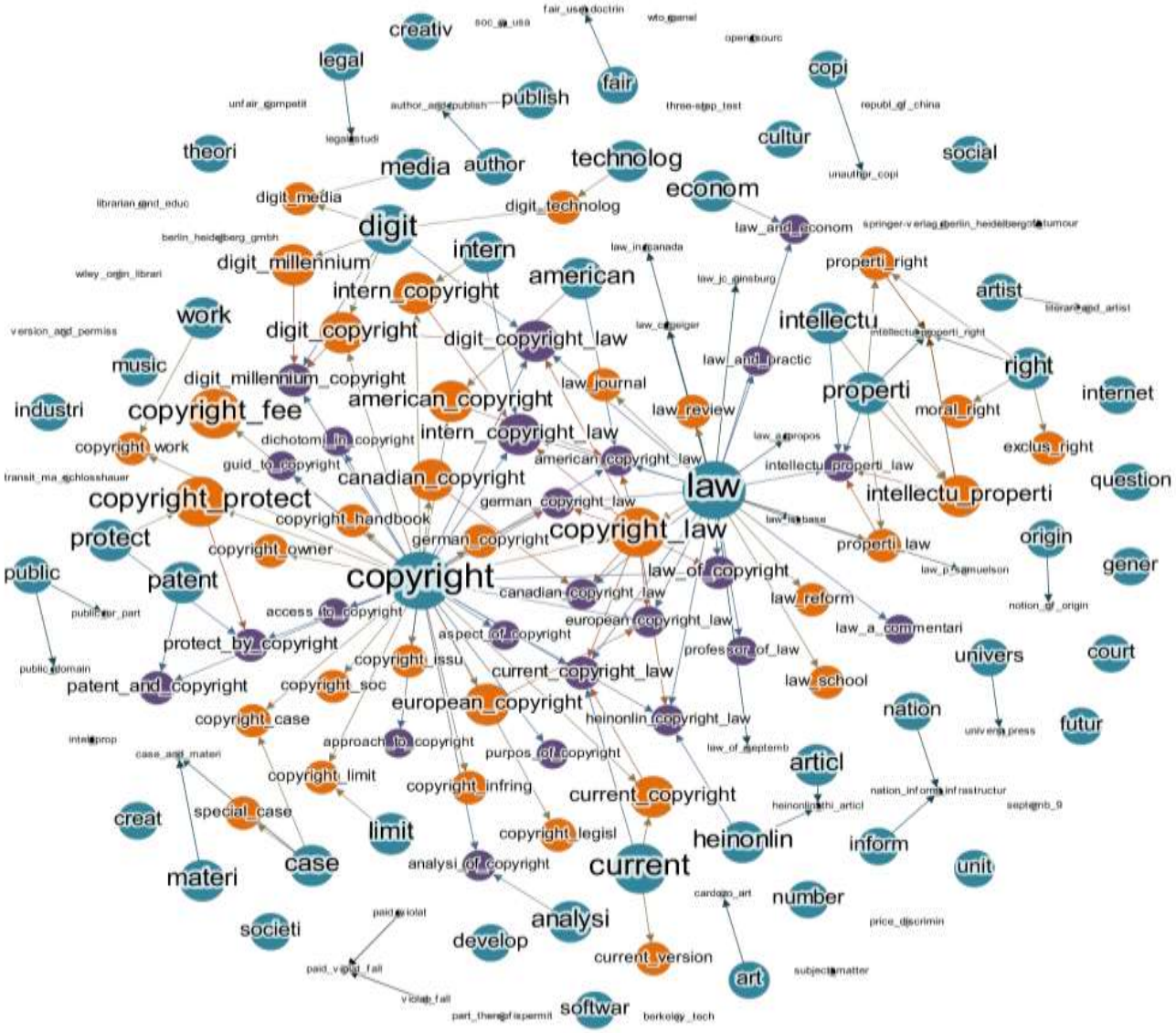

Рис. 4. Мережа природніх ієрархій термінів розміром 50+50+50
для предметної області “Copyright Law”

Також за допомогою засобів програмного забезпечення Gephi були отримані такі параметри створеної мережі: кількість вузлів – 150; кількість зв’язків – 144; щільність мережі – 0.006; кількість зв’язаних компонент – 48; середня довжина шляху – 1; середній коефіцієнт кластеризації – 0.068.

Висновки.

У статті розглянуто методику створення мережі зі слів та словосполучень – алгоритм формування мереж природніх ієрархій термінів для масиву тематично пов’язаних текстових документів. Розглянута методика була застосована для створення

моделей предметних областей “Criminal Law” та “Copyright Law”. На основі найбільшої вільно-доступної пошукової системи, яка індексує повний текст наукових публікацій – Google Scholar, були попередньо підготовлені текстові корпуси за запитами “Criminal Law” та “Copyright Law” об’ємом 385 документів та 490 документи відповідно. Були отримані мережі природніх ієрархій термінів для масиву текстових документів тематично пов’язаних з “Criminal Law” та “Copyright Law”.

В якості допоміжних інструментів для дослідження були використаний пакет візуалізації та моделювання графів Gephi (http://gephi.org) та власний набір спеціально розроблених модулів на мові програмування Python. Було встановлено, що створені мережі за топологічною особливістю мають малий середній коефіцієнт кластеризації. Невелика середня довжина шляху підтверджує припущення про те, що ця мережа є “малим світом” (Small World).

Отже, мережа природної ієрархії термінів, що створюється повністю автоматично, може розглядатися як основа для подальшого автоматизованого формування термінологічних онтологій за участю експертів. Розглянута у статті методика створення направленої мережі зі слів та словосполучень сприятиме формуванню й удосконаленню понятійного і термінологічного апарату у правовій сфері та гармонізації національного і міжнародного права. Також результати дослідження можуть бути використані під час створення персональних пошукових інтерфейсів користувачів інформаційно-пошукових систем, що, в свою чергу, дозволить спростити процес пошуку необхідної інформації.

Використана література

1. Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Отбор словосочетаний для словаря системы автоматической обработки текстов. *Компьютерная лингвистика и интеллектуальные технологии*: труды международной конференции “Диалог – 2008”. Москва, 2008. С. 339-344.

2. Филиппович Ю.Н., Прохоров А.В. Семантика информационных технологий: опыты словарно-тезаурусного описания. Москва: МГУП, 2002. – 368 с.

3. Ланде Д.В. Елементи комп’ютерної лінгвістики в правовій інформатиці. Київ: НДІІП НАПрН України, 2014. 168 с.

4. Lande D.V., Snarskii A.A., Yagunova E.V., and Pronoza E. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. In: *Proceedings of the 12th Mexican International Conference on Artificial Intelligence*, 2013. Pp. 209-215.

5. Yagunova E.D. and Lande D.V. Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects. CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference *“Digital libraries: Advanced Methods and Technologies, Digital Collections”*. Pereslavl-Zalessky. Russia, 2012. Pp. 150-159.

6. Wang M., Xu H., Tian L. and Stanley H.E. Degree distributions and motif profiles of limited penetrable horizontal visibility graphs. *Physica A: Statistical Mechanics and its Applications*. 2018.

7. Wang M., Vilela A.L., Du R., Zhao L., Dong G., Tian L., and Stanley H.E. Exact results of the limited penetrable horizontal visibility graph associated to random time series and its application. *Scientific reports.* 8(1). 2018.

8. Wang M., Vilela A.L., Du R., Zhao L., Dong G., Tian L., and Stanley H.E. Topological properties of the limited penetrable horizontal visibility graph family. *Physical Review E*. 97(5), 2018.

9. Luque B., Lacasa L., Ballesteros F., and Luque J. Horizontal visibility graphs: Exact results for random time series. *Physical Review E.* 80(4). 2009.

10. Gutin G., Mansour T., and Severini S. A characterization of horizontal visibility graphs and combinatoris on words. *Physica A.* 390. 2011. Pp. 2421-2428.

11. Lande D.V. Building of Networks of Natural Hierarchies of Terms Based on Analysis of TextsCorpora. E-preprint ArXiv 1405.6068.

12. Manning C.D., Raghavan P., and Schütze H. An Introduction to Information Retrieval. *Cambridge University Press.* 2009. Pp. 22-36.

13. Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. № 24(5). 1998. Pp. 513-523.

14. Ullman J.D. Data Mining, Mining of massive datasets. *Cambridge University Press*. 2011. Pp. 1-17.

15. Beel J., GIPP B., Langer S., Breitinger C. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*. 17(4). 2016. Pp. 305-338.

16. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, MCB University Press*. 60. 2004. Pp. 493-502.

17. Lande D.V., Dmytrenko O.O., Snarskii A.A. Transformation texts into complex network with applying visability graphs algorithms. *Информационные технологии и безопасность:* материалы XVIII Международной научно-практической конференции ИТБ-2018. Киев: ООО "Инжиниринг". 2018. С. 20-33. CEUR Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn: de:0074-2318-4. Selected Papers of the XVIIIШ International Scientific and Practical Conference on Information Technologies and Security (ITS 2018).

18. Jongejan B., and Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In the *Proceeding of the ACL-2009*. Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Singapore. August 2-7, 2009. Pp. 145-153.

19. Baeza-Yates R., Ribeiro B. D. A. N. Modern information retrieval. New York: ACM Press. Harlow. England: Addison-Wesle. 2011.

20. Porter M.F. An algorithm for suffix stripping. *Program.* Vol. 14. No. 3. 1980. Pp. 130-137,

21. Willett P. The Porter stemming algorithm: then and now. *Program: Electronic Library and Information Systems*. Vol. 40. No 3. 2006. Pp. 219-223.

22. Lande D.V., and Snarskii A.A. Compactified HVG for the Language Network. In: *Proceedings of the International Conference on Intelligent Information Systems*: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science. 20-23 Aug. 2013, Chisinau, Moldova: Proceedings IIS, Institute of Mathematics and Computer Science. 2013. Pp. 108-113.

23. Lande D.V., Snarskii A.A., and Yagunova E.V. Application of the CHVG-algorithm for scientific texts. In: *Proceedings of the Open Semantic Technologies for Intelligent Systems (OSTIS)*, February 20 – 22th. Minsk. 2014. Pp. 199-204.

24. Kleinberg J.M. Authoritative sources in a hyperlink environment. *Journal of the ACM JACM*. 46 (5). 1999. Pp. 604-632.

25. Langville A.N., and Meyer C.D. Google's PageRank and beyond: the science of searchengine rankings. Princeton university press. 2011.

26. Cherven K. Network Graph Analysis and Visualization with Gephi. Packt Publishing, 2013.

27. Kleinberg J. Navigation in a small world. *Nature*. 2000. 406 (6798). P. 845.

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~