2. Батракова Т. І. Особливості та принципи цифрової економіки в Україні /Т. Ш Батракова, В. Ю. Линовецька // Економічні студії. – 2018. – №2. – С. 94-96.

3. Пуцентейло П. Р. Цифрова економіка як новітній вектор реконструкції традиційної економіки / П. Р. Пуцентейло, О. О. Гуменюк // Інноваційна економіка. – 2018. – № 5-6. – С. 131-143.

4. Танскотт Д. Электронно-цифровое общество. Плюсы и минусы сетевого интеллекта / Д. Транскотт. – К.: INT Пресс: М.: Рефл. Бук, 1999. – 432 с.

5. Coase, R. H. (1937), The Nature of the Firm. Economica, 4: 386-405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x

6. Головенчик, Г. Г. Цифровая экономика / Г. Г. Головенчик, М. М. Ковалев. – Минск: Изд. центр БГУ, 2019. – 395 с.

7. Булдыгин С. С. Концепция промышленной революции: от появления до наших дней. Вестник Томского государственного университета. – 2017. – № 420. – С. 91–95. <https://cyberleninka.ru/article/n/kontseptsiya-promyshlennoy-revolyutsii-ot-poyavleniya-do-nashih-dney>

8. Баранов О. А. Економіка результату та право / «Економічні свободи та інституції: правове регулювання та ефективність», збір. мат.між. науково-практич. конф. (22-23 жовтня 2020 р., м. Ужгород). – Ужгород: РІК-У, 2020. С. 23-29.

Ланде Д. В.
доктор технічних наук, професор,
завідувач відділу спеціалізованих засобів
моделювання Інституту проблем
реєстрації інформації НАН України
Дмитренко О. О.
аспірант Інституту проблем
реєстрації інформації НАН України

ФОРМАЛІЗАЦІЯ ЗНАНЬ ТА ПОБУДОВА ТЕРМІНОЛОГІЧНИХ ОНТОЛОГІЙ У ПРАВОВІЙ ГАЛУЗІ

Сучасні інформаційно-комунікаційні технології та загалом інформаційний простір розвиваються швидше, ніж коли-небуть раніше. Такий процес характеризується відповідно стрімким збільшенням об’ємів даних [1], які продукуються елементами інформаційного простору, зокрема, документами та найрізноманітнішими джерелами даних – файлами,

електронними листами, веб сторінками та іншими джерелами не залежно від форматів їх подання. Важливо зазначити й той факт, що обсяг вищезгаданих даних подвоюється приблизно кожні 18 місяців [2]. Унаслідок цього за п'ять попередніх років людством було вироблено інформації більше, ніж за всю попередню історію. Та такий інформаційний сплеск, або так званий інформаційний вибух, супроводжується не лише припливом нових цінних знань. Основну частину накопичених даних складають неструктуровані дані (близько 95%), в тому числі й непотрібні та шумові, і лише зовсім мала частина представляє собою певну інформацію, яка може бути використана під час прийняття рішень.

Тож в результаті критичної невідповідності між розвитком сучасних інформаційних систем і збільшенням динамічних інформаційних потоків у глобальних комп'ютерних мережах перед інформаційним суспільством постає ряд проблем. Одна із них 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 та накопичуються.

Тож величезні об'єми інформаційних потоків та динамічних текстових даних, що накопичуються у глобальних комп'ютерних мережах обумовлюють актуальність процесу концептуалізації цих даних та їх подальшої формалізації у вигляді певної онтологічної моделі [3].

Оскільки науково-технічний прогрес вплинув і на правову галузь, то кількість нормативно-правових документів поданих у електронній формі, а отже, і кількість інформації, з якою доводиться мати справу експерту у цій сфері, теж постійно зростає. І для прийняття законних рішень інколи необхідно ознайомлюватися з тисячами документів, свідомо відкидаючи інформаційний шум. Тож отримання коротких і водночас найважливіших відомостей чи викладок з одного або з декількох текстових документів, у вигляді так званих рефератів, а також генерація лаконічних інформаційно-насичених звітів на основі коротких анотацій або дайджестів є актуальним заданням. А отже, проблема комп'ютеризованої обробки правової інформації та удосконалення або розробка нових систем автоматичного реферування, які могли б прийнятною продуктивністю і якістю обробляти великі об'єми правових документів й надавати спрощений доступ до головного змісту правового тексту, виокремлювали з нього найважливіші відомості чи викладки, ідеї та заздалегідь заявлені змістові аспекти без необхідності опрацьовувати великий за обсягом текстовий документ або текстовий корпус є також актуальною. Також не менш важливим є завдання

виявлення дублюючої інформації та протиріч у нормативно-правових документах.

Визначальною особливістю правової інформації є те, що пов'язані з нею тексти не у повній мірі вільнодоступні та неструктуровані. Наявна структура окремих видів документів і застосування найкращих універсальних систем реферування [4, 5] не дає задовільних результатів. Це важливо враховувати під час вибору потрібного методу чи підходу для вирішення вищезгаданої проблеми у галузі права.

Той факт, що під час комп'ютеризованої обробки текстових даних багато задач лежать на перетині між математичними науками та лінгвістикою, відкриває широкі можливості для застосування потужного математичного апарату (такого, як теорія графів та складних мереж) та лінгвістичної теорії (що враховує семантичну та синтаксичну структуру тексту).

У цій роботі для побудови термінологічної онтології, придатної для автоматизованої обробки, застосовується лінгвомережева модель представлення текстових даних. Одним із видів такої мережевої моделі є мережа, що побудована із ключових слів та словосполучень (або просто – мережа термінів). В ній вузли відповідають окремим ключовим поняттям предметної галузі, а ребра – семантико-семантичним зв'язкам між ними.

Для виокремлення ключових термінів застосовується комп'ютерна обробка природномовних текстів, що включає автоматичну сегментацію на окремі речення, розбиття на токени та розмічування частин мови й присвоєння тегів кожному слову (Part-of-Speech tagging) [6]. Оскільки запропоновані у цій роботі методи орієнтовані на роботу з англомовними правовими текстами, то відповідно використовується класичний набір тегів, що сформований на основі Brown Corpus (стандартний корпус університету Брауна) та має назву »The Penn Treebank" [7].

Використовуючи шаблони ключових слів та словосполучень, що представлені у роботі [6], формується послідовність термінів. Далі здійснюється видалення одиничних стоп-слів (окремих артиклів, прийменників, сполучників, деяких дієслів, прислівників та займенників), які не несуть ніякого інформативного навантаження.

На наступному етапі для кожного сформованого терміна у порядку його зустрічання у тексті формується так званий кортеж. Кожен елемент кортежу складається з трьох значень: перше – термін (слово або словосполучення, що отримане за одним із шаблонів); наступне – тег, який присвоюється слову в залежності від його приналежності до певної частини мови; останній елемент такого набору – числове значення GTF [8] (глобальна частота терміна, що використовується для статистичного

зважування слів та словосполучень, що входять у сформовану на попередньому етапі послідовність). Важливо зазначити, що GTF обчислюється з урахуванням двох попередніх значень кортежу – терміна та частини мови, до якої він належить. Кількість таких однакових кортежів у всьому тексті, що нормована на загальну кількість сформованих термінів, і визначає значення третього елемента.

Для встановлення ненаправлених зв'язків між ключовими термінами в межах кожного окремого речення застосовується алгоритм графа горизонтальної видимості для часових рядів (Horizontal Visibility Graph algorithm – HVG) [9]. Сформована на попередньому етапі послідовність числових значень GTF, які відповідають окремим кортежам, є тим часовим рядом, який завдяки алгоритму HVG трансформується у ненаправлену мережу.

Для встановлення напрямків зв'язків враховувались правила, представлені у роботі [6]. Після об'єднання однакових вузлів сумарна кількість однаково-направлених зв'язків між цими вузлами визначала вагове значення зв'язку.

Для апробації представленої у цій роботі методики побудови мережі термінів було використано вільнодоступний правовий документ «Convention on the Rights of the Child», поданий англійською мовою [10]. В результаті було отримано онтологічну модель у вигляді мережі із ключових термінів (рис. 1).

Література

1. Mayer-Schцnberger V., Cukier K. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013.

2. Humanity Doubles Its Data Creation Every 18 Months, And It Has Powerful Implications. URL: https://www.fluxmagazine.com/data-creation-powerful-implications/ (дата звернення: 21.03.2021).

3. Lande D. V., Radziievska O. H. 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. CEUR Workshop Proceedings (ceur-ws.org). 2020. Vol-2604. pp. 32-43.

4. Best Text Summarizing Tool for Academic Writing [For Free]. URL: https://ivypanda.com/online-text-summarizer (дата звернення: 21.03.2021).

5. Ланде Д. В., Яньцін Чжао, Моцзі Вей, Шівей Чжу, Цзяньпін Ґо Система анотування китайської правової інформації // Інформація і право. 2018. № 3(26). С. 66-71.

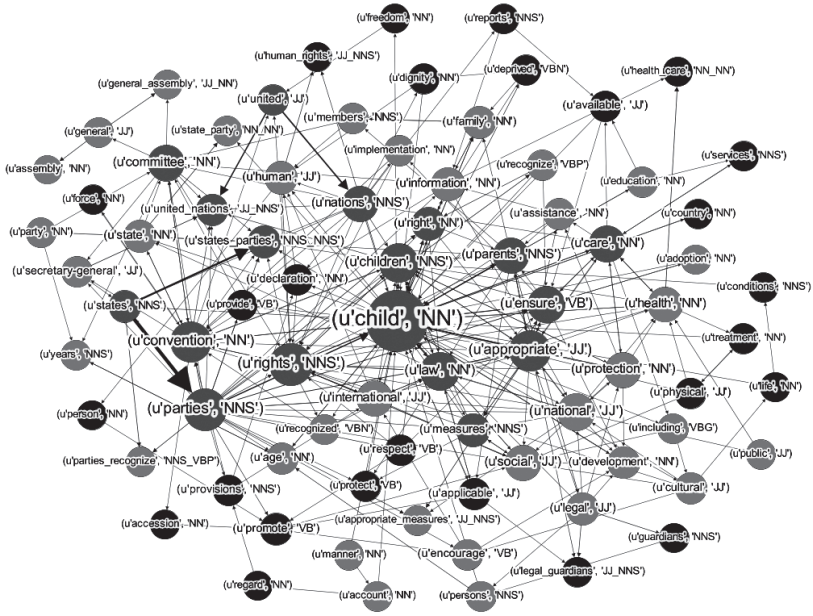

Рис. 1 Мережа термінів, що відповідає правовому документу «Convention on the Rights of the Child»

6. Дмитренко О. О. Побудова направлених зважених мереж термінів із застосуванням Part-of-speech tagging. // Реєстрація, зберігання і обробка даних, 2020. Т. 22, № 4. С. 47-55. DOI: 10.35681/1560-9189.2020.22.4.225914.

7. Marcus M., Santorini B., Marcinkiewicz M. A. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics. 1993. Volume 19 № 2. pp. 313-330.

8. D. Lande, O. Dmytrenko, O. Radziievska, Determining the Directions of Links in Undirected Networks of Terms, in: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference «Information Technologies and Security» (ITS 2019), volume 2577, 2019, pp. 132–145. ISSN 1613-0073.

9. Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs: Exact results for random time series. Physical Review E. 2009. Volume 80. № 4. 046103.

10. International Covenant on Economic, Social and Cultural Rights. URL: https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx (дата звернення: 22.03.2021).