УДК 004.67

ЛАНДЕ Д.В., *доктор технічних наук,*
Інститут проблем реєстрації інформації НАН України

ФОРМУВАННЯ МЕРЕЖ ПРИРОДНИХ ІЄРАРХІЙ ТЕРМІНІВ НА ОСНОВІ АНАЛІЗУ ТЕКСТОВИХ КОРПУСІВ З ПРАВОВОЇ ТЕМАТИКИ

Анотація. *Пропонується методика виявлення і побудови мереж ієрархій термінів на основі аналізу текстових корпусів відповідної тематики. Методика базується на застосуванні методології компактифікованих графів горизонтальної видимості. Побудовано і досліджено мережі понять, сформовані на основі даних моделі Електронної енциклопедії законодавства України, а також тематичного фрагменту бази даних “Україніка наукова”.*

Ключові слова: *мовна мережа, ієрархія термінів, правова інформація, граф видимості, візуалізація.*

Аннотация. *Предлагается методика выявления и построения сетей иерархий терминов на основе анализа текстовых корпусов соответствующей тематики. Методика базируется на применении методологии компактифицированных графов горизонтальной видимости. Построены и исследованы сети понятий, сформированные на основе данных модели Электронной энциклопедии законодательства Украины, а также тематического фрагмента базы данных “Украиника научная”.*

Ключевые слова: *сеть языка, иерархия терминов, правовая информация, граф видимости, визуализация.*

Summary. *The methods of identifying and building of hierarchies of terms networks based on the analysis of text corps on relevant topics. The procedure is based on the application of the methodology of horizontal visibility graphs. Constructed and investigated concept networks are formed based on models of electronic encyclopedia of ukrainian legislation, as well as fragment of thematic database “Ukrainika Naukova”.****Keywords:*** *language network, hierarchy of terms, legal information, visibility graph, visualization.*

Постановка проблеми. На цей час актуальними є задачі побудови онтологій з визначених галузей знань, зокрема, з правової тематики. Зрозуміло, побудова великих галузевих онтологій – це складна проблема, яка потребує великих ресурсних витрат. У будь-якому разі, певним етапом побудови загальної онтології є побудова відповідних тезаурусів, термінологічних онтологій [1].

В цій роботі надається методика побудови мережі природної ієрархії термінів, яку можна розглядати як “квазіонтологію”, основу для формування відповідної термінологічної онтології. Мережа природної ієрархії термінів базується на інформаційно-значущих елементах тексту, опорних словах і словосполученнях, методологію виявлення яких наведено в роботі [2]. Використання таких елементів дозволяє формувати пошукові образи, зокрема, при обробці правової інформації, виявляти такі компоненти тексту, як колокації, надфразові єдності [3], охоплювати цілі галузі знань як основи для подальшої побудови загальних онтологій [4].

Опірні слова і словосполучення для побудови природних ієрархій термінів вибираються з урахуванням такої властивості слів, як “розпізнавальна” або дискримінантна сила [5], яка має важливе значення при аналізі текстів з правової тематики, зокрема, при вирішенні завдання формування електронної енциклопедії на основі аналізу всього масиву законодавчих актів України.

Разом з тим, однієї цієї властивості виявляється недостатньо при побудові тезаурусів і онтологій. Інколи слова з низькою дискримінантною силою, зокрема, найчастіші слова з вибраної предметної області (наприклад, слова “закон”, “постанова”, “держава”, “Україна” і т. д.) виявляються найважливішими для задачі, що розглядається.

Метою статті є опис і практичне обґрунтування методики формування і візуалізації мережі природних ієрархій термінів (далі – МПІТ) на основі аналізу текстових корпусів правової спрямованості. “Природність” ієрархій термінів у цьому випадку розуміється як відмова при її формуванні від спеціальних синтетичних методик (наприклад, семантичного аналізу), усі зв’язки в такій мережі визначаються природним застосуванням слів і словосполучень, що екстрагуються із текстових корпусів статистично значущих обсягів. Мережа природних ієрархій термінів, що має формуватися повністю автоматично, може розглядатися як основа для подальшого автоматизованого формування термінологічної онтології, яка описує різні терміни (слова і сталі словосполучення), що відповідають поняттям цільових предметних галузей, а також правила вибору термінів, які відповідають екземплярам заданих понять.

Виклад основних положень. Запропонований автором алгоритм формування мережі природних ієрархій, що розглядається в цій роботі, передбачає реалізацію послідовності кроків, що охоплюють попередню обробку вихідного текстового корпусу, визначення і сортування термінів, вибір необхідної кількості найбільш вагомих (найбільших вузлів компактивікованого графу горизонтальної видимості), побудову МПІТ і її відображення. Розглянемо ці кроки детально.

1. Як вхідні текстові корпуси з правової тематики нами розглядаються два корпуси, а саме: тестове наповнення Електронної енциклопедії законодавства України [6] (на цей час створено діючу модель цієї енциклопедії, яка містить понад 6000 статей, доповнених як зовнішніми посиланнями на законодавчі акти, кодекси України, так і внутрішніми – посиланнями з одних статей на інші) і тематичний фрагмент реферативної бази даних “Україніка наукова”, яка входить до системи реферування української наукової літератури [7]. В ній містяться реферати статей наукових періодичних видань та збірників, монографій, праць наукових конференцій, авторефератів дисертацій, довідників та словників, підручників, що видаються в Україні. Аналізувалися дані, що знаходились у базі даних за станом на листопад 2013 року, що становило близько 500 000 записів, серед яких до правової тематики у відповідності з рубрикатором Національної бібліотеки України ім. Вернадського відносяться 28600 документів з рубрики “Держава і право. Юридичні науки”. Для проведення досліджень застосовуються засоби фільтрації та аналізу даних, які забезпечують виділення тематичних фрагментів вихідної бази даних, окремих записів і полів.

Попередня обробка цих текстових корпусів передбачає здійснення стемінгу (вилучення флексій) слів, що входять до цих корпусів, вилучення нетекстових символів, а також чітке виділення окремих текстових частин корпусу, які трактуються як окремі документи для подальшої обробки.

2. На другому етапі кожному окремому слову з текстового корпусу, що аналізується, ставиться у відповідність оцінка його “дискримінантної сили”, а саме TFIDF (у канонічному виді, рівну добутку частоти слова у фрагменті тексту (Term Frequency) на двійковий логарифм від величини, зворотної кількості фрагментів тексту, в яких це слово зустрілось, – Inverse Document Frequency) [8]. Для цього для кожного слова i, що входить до текстового корпусу, який складається з N документів,

підраховується кількість документів $df(i)$, в яких міститься це слово, а також загальна частота входження даного слова i у текстовий корпус – $n(i)$. Після цього розраховується середнє значення TFIDF вагової оцінки для кожного слова за формулою:

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{df(i)}\right).$$

3. Виконується те ж саме, що і на попередньому кроці, тільки для словосполучень із двох слів (біграм).

4. Виконується те ж саме, що і на попередньому кроці, тільки для словосполучень із трьох слів (триграм).

5. Для послідовностей термінів і їх вагових значень за TFIDF будуються компактифіковані графи горизонтальної видимості (CHVG) [9] і виконується повторне визначення вагових значень слів за цим алгоритмом. При цьому ряди з цифрових значень, відповідних термінам, перетворяться в графи горизонтальної видимості, в яких вузлам відповідають не лише цифрові значення, але самі слова, що виражають певне змістовне значення. Ця процедура дозволяє враховувати у подальшому, крім термінів з великою дискримінантною силою, також терміни, що мають велике значення для загальної тематики текстового корпусу, але відрізняються великою частотою появи.

Після цього всі терміни сортуються за зменшенням розрахованих вагових значень відповідних вузлів CHVG. Подальшому аналізу, крім того, не підлягають терміни, що входять до так званого стоп-словника. Це, як правило, фіксований набір службових слів, які не відіграють суттєвої ролі для інформаційної структури текстів.

6. Експертним методом визначається необхідний обсяг МПІТ (число N), після чого обирається відповідна кількість одиничних слів, біграм і триграм (всього $N \times N \times N$ елементів) з найбільшими ваговими значеннями за CHVG.

7. З відібраних на попередньому кроці елементів будується мережа природних ієрархій термінів, в якій як вузли розглядаються самі терміни, а зв’язки відповідають входженням одних термінів до других.

На Рис. 1 проілюстровано принцип побудови зв’язків МПІТ. Окремі геометричні фігури на цій ілюстрації відповідають одиничним словам. Першому рядку відповідає вибрана множина одиничних слів, другому – множина біграм, а третьому – множина триграм. Якщо одиничне слово входить до біграми або триграми, або біграма входить до триграми, утворюється зв’язок, який позначається стрілкою. Множина вузлів, яким відповідають терміни, і зв’язків і утворює трирівневу мережу природної ієрархії термінів.

8. На останньому етапі формування МПІТ здійснюється її відображення за допомогою програмного пакету аналізу і візуалізації складних мереж Gephi (https://gephi.org/). Для завантаження мереж природних ієрархій термінів до баз даних цієї системи достатньо привести відповідну матрицю інцідентності до загальноприйнятого формату csv.

Для побудованих мереж природних ієрархій термінів за зазначеними текстовими корпусами було визначено розподіл вихідних степенів вузлів, який виявився близьким до степеневого ($p(k) = Ck^{\alpha}$), тобто ці мережі є безмасштабними. Були проведені розрахунки параметрів цих мереж. У результаті виявилось, що коефіцієнт α для них складає близько 2,15.

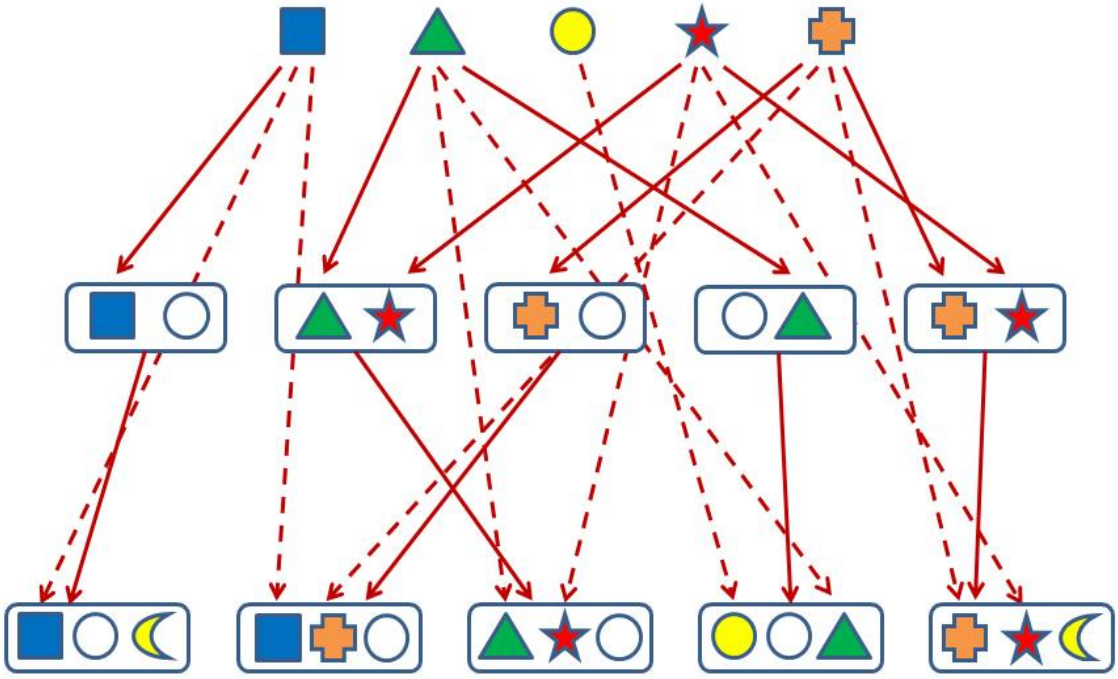

Рис. 1 – Формування зв’язків у трирівневій мережі природної ієрархії термінів

На Рис. 2 представлено загальний вигляд мережі природній ієрархії термінів розміром 50×50×50, яку візуалізовано засобами системи Gephi.

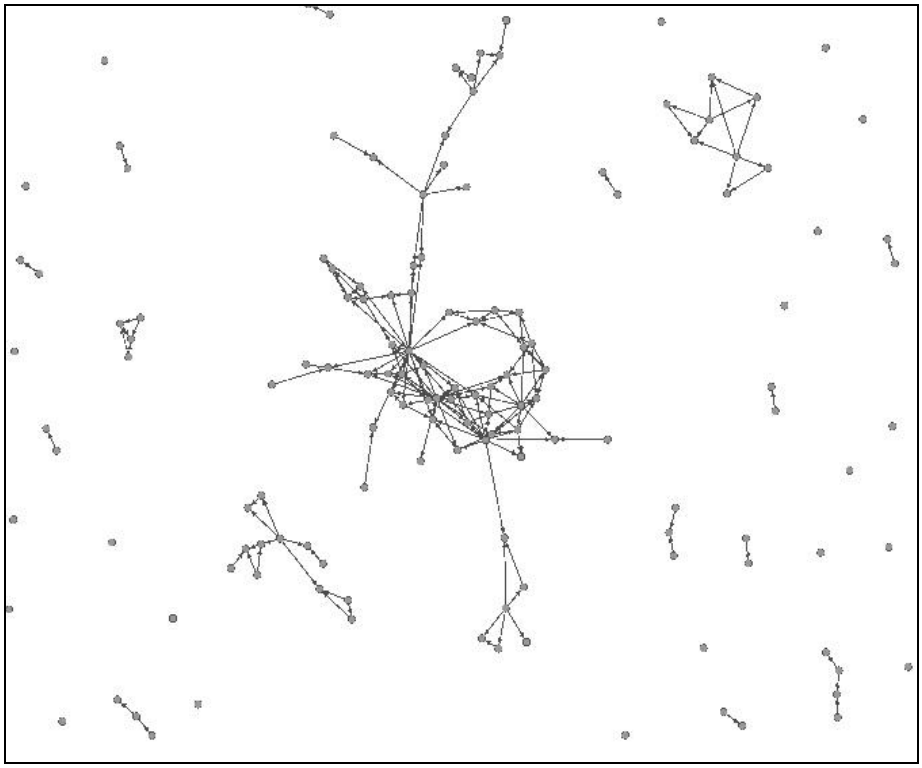

Рис. 2 – Загальний вигляд МПІТ розміром 50×50×50

На Рис. 3 наведено окремі фрагменти мережі природної ієрархії термінів, що відповідають вибраним поняттям.

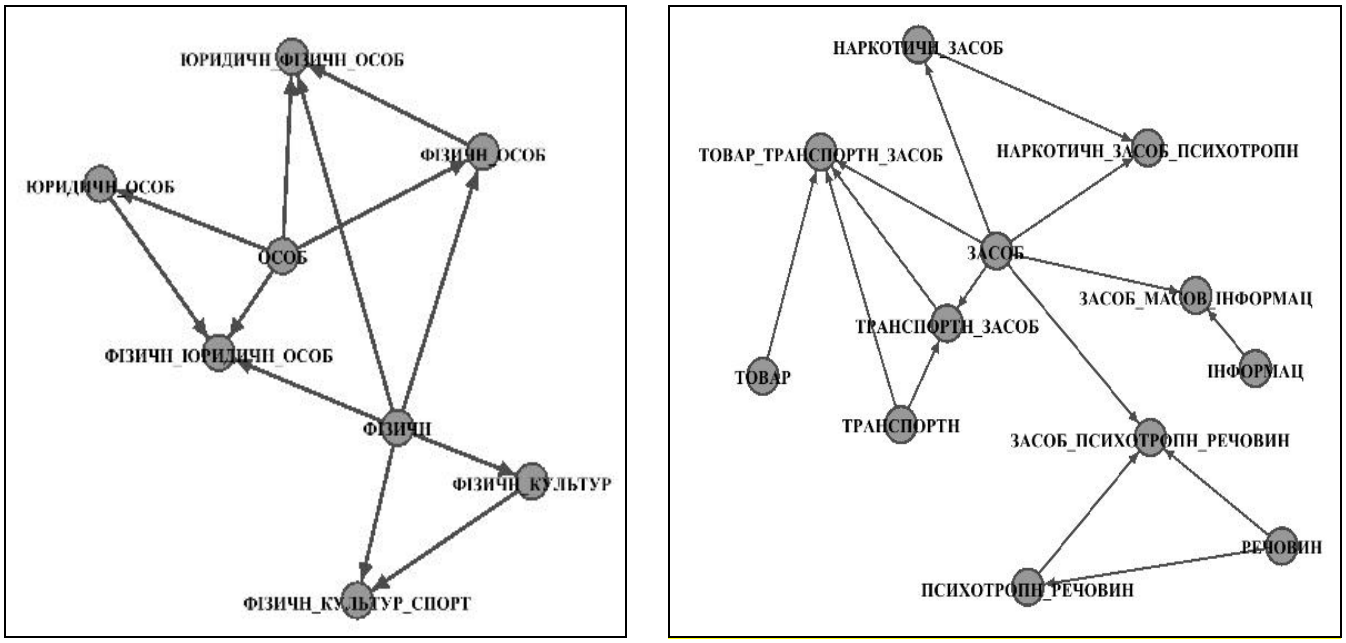

Рис. 3 – Фрагменти МПІТ, побудованої за вибраними текстовими корпусами

Висновки

У результаті проведених досліджень:

- Запропоновано алгоритм виявлення і побудови природних ієрархій термінів на базі аналізу текстових корпусів.
- На основі запропонованого алгоритму за двома текстовими корпусами побудовано природні ієрархії термінів.
- Досліджено властивості мережі природних ієрархій термінів, яка виявилась скейл-фрі (безмасштабною) за вихідними зв’язками.
- Запропоновані наочні засоби візуалізації мережі природних ієрархій термінів.
- Мовні мережі, побудовані за допомогою запропонованої методики, можна розглядати як базу для побудови загальної онтології з правової тематики, а також використовувати на практиці як готові до застосування засоби навігації у базах даних відповідної тематики, а також для контекстної підказки користувачам відповідних інформаційно-пошукових систем.

Використана література

1. Ланде Д.В. Елементи комп’ютерної лінгвістики в правовій інформатиці / Д.В. Ланде. – К. : НДІІП НАПрН України, 2014. – 168 с.

2. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text / 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209-215.

3. Ягунова Е.В., Ландэ Д.В. Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов ; труды 14-й Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” – RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. – С. 196-205.

4. Харламов А.А. Раевский В.В. Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды // Речевые технологи. – 2008. – № 3.– С. 27-35.

5. Ланде Д.В. Методи оцінки рівня дискримінантної сили слів у текстах з правової тематики // Правова інформатика. – 2012. – № 3 (35). – С. 5-9.

6. Ланде Д.В., Брайчевський С.М. Можливості довідкових мережевих ресурсів для створення електронної енциклопедії законодавства України // Інформація і право. – № 2(38)/2013. – С. 72-76.

7. Ланде Д.В., Балагура І.В. Дослідження мереж співавторства у правовій науці по базі даних “Україніка наукова” // Правова інформатика. – № 4(36)/2012.. – С. 50-57.

8. Salton G., McGill M.J. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.

9. Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs : Exact results for random time series // Physical Review E, 2009. – P. 046103-1 – 046103-11.

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