Науково-дослідний інститут інформатики і права
Національної академії правових наук України
Інститут законодавства Верховної Ради України

ПРАВОВА ІНФОРМАТИКА

№ 2(38) / 2013

Заснований у грудні 2003 року	**НАУКОВИЙ ФАХОВИЙ ЖУРНАЛ З ПИТАНЬ ІНФОРМАТИКИ, ІНФОРМАТИЗАЦІЇ ІНФОРМАЦІЙНОГО ПРАВА ТА ІНФОРМАЦІЙНОЇ БЕЗПЕКИ**	Видається щоквартально

Свідоцтво про державну реєстрацію журналу: КВ № 8254 від 22.12.03 р., видане Державним комітетом телебачення і радіомовлення України

У журналі можуть публікуватися матеріали стосовно дисертаційних робіт на здобуття наукових ступенів доктора і кандидата юридичних наук (Постанова президії ВАК України від 08.07.09 р. № 1-05/3) **та технічних наук** (Постанова президії ВАК України від 10.02.10 р. № 1-05/1)

Видавець журналу: © Науково-дослідний інститут інформатики і права Національної академії правових наук України

Адреса редакції:
01032, м. Київ, вул. Саксаганського, 110-В.
Тел.: 234-94-56, 234-91-33

Створення оригінал-макета, дизайн та наукове редагування – Брижко В.М.
Редагування – Майстренко І.А. (укр., англ.).
Формат 70 х 108/16. Папір офсетний. Гарнітура Times.
Офсетний друк. Ум. друк. арк. 8.75. Тираж 100 прим.
Виготовлено з оригінал-макета в друкарні ТОВ “ПанТот”, м. Київ, вул. Щорса, 29.

З М І С Т

Інформатика, інформатизація

Інформаційне право та інформаційна безпека

З інших юридичних наук

До питання якості наукових досліджень

Рекомендовано до друку Вченою радою НДІІП НАПрН України, протокол № 5 від 20.05.13 р.

УДК 004.67

ЛАНДЕ Д.В., *доктор технічних наук,*
СНАРСЬКИЙ А.О., *доктор фізико-математичних наук, професор*

ГРАФИ ГОРИЗОНТАЛЬНОЇ ВИДИМОСТІ ЯК ЗАСІБ ВИТЯГУ ІНФОРМАЦІЙНО-ЗАНЧУЩИХ СЛІВ ІЗ ЗАКОНОДАВЧИХ АКТІВ

Анотація. *Пропонується методика створення і використання компактифікованих графів горизонтальної видимості для текстів законодавчих актів з метою виявлення слів, які визначають їх інформаційну структуру. Показано, що такі графи є безмасштабними, а також, що серед вузлів з найбільшими степенями є слова, що визначають як структуру зв'язності тексту, так і його інформаційну структуру.*

Ключові слова: *мережа мови, складна мережа, безмасштабна мережа, граф видимості.*

Аннотация. *Предлагается методика создания и использования компактифицированных графов горизонтальной видимости для текстов законодательных актов с целью выявления тех слов, которые определяют их информационную структуру. Показано, что такие сети являются безмасштабными, а также, что среди узлов с наибольшими степенями имеются слова, определяющие как структуру связности текста, так и его информационную структуру.*

Ключевые слова: *сеть слов, сложная сеть, безмасштабная сеть, граф видимости.*

Summary. *The methods of creation and use of compactifying horizontal visibility graph are offered for texts of legislative acts with the purpose of exposure of those words which determine their informative structure. It was found that the networks constructed in such way are scale free, and have a property that among the nodes with largest degrees there are words that determine not only a text structure communication, but also its informational structure.*

Keywords: *language network, complex network, scale-free network, visibility graph.*

Постановка проблеми. На даний час актуальним є завдання визначення того, які з важливих структурних елементів тексту виявляються інформаційно-значущими, такими, що визначають інформаційну структуру тексту. Використання таких елементів як опорних слів дозволяє формувати онтології, тезауруси, пошукові образи, зокрема, при обробці законодавчих актів та іншої нормативно-правової інформації. Такі елементи можуть, зокрема, використовуватися також для ідентифікації таких компонент тексту, як коллокации, надфразова єдність [1].

Опірні слова для пошуку в тексті та автоматичного екстрагування значущих фрагментів вибираються з урахуванням такої властивості слів, як “розпізнавальна” або дискримінантна сила. При аналізі текстів з правової тематики, зокрема, при вирішенні завдання формування електронної енциклопедії на основі аналізу всього масиву законодавчих актів України, оцінка дискримінантної сили окремих слів має найважливіше значення [2].

Метою статті є опис і практичне обгрунтування методики виявлення опірних слів за допомогою так званних мереж мови (Language Network), що пропонується аторами. Разом з послідовним аналізом текстів, побудова мереж, вузлами яких є їх елементи – слова або словосполучення, фрагменти природної мови, дозволяє виявляти структурні елементи тексту, без яких він втрачає свою зв’язність. Відомо декілька підходів до побудови мереж з текстів, так званих мереж мови, і різні способи інтерпретації вузлів і зв’язків, що приводить, відповідно, до різних видів представлення

таких мереж. Вузли можуть бути сполучені між собою, якщо відповідні їм слова стоять поряд у тексті [3, 4], належать до одного речення або абзацу [5], сполучені синтаксично [6, 7] або семантично [8, 9].

Виклад основних положень. У рамках концепції складних мереж (Complex Networks) [10, 11] запропоновано декілька методів побудови мереж на основі часових рядів, серед яких можна назвати декілька методів побудови графів видимості [12], зокрема, так званий граф горизонтальної видимості (Horizontal Visibility Graph – HVG) [13, 14]. Ці підходи також дозволяють будувати мережеві структури на підставі текстів, в яких окремим словам або словосполученням деяким спеціальним чином поставлені у відповідність числові вагові значення. Як функція, що ставить у відповідність слову число, можна розглядати, наприклад, порядковий номер унікального слова у тексті, довжину слова, загальноприйняту оцінку 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) або її варіанти [15, 16], а також інші вагові оцінки.

Для підрахунку вагової оцінки TFIDF з повного тексту, що складається з N слів, текст розбивається на фрагменти, що містять задану кількість слів M (наприклад, $M = 500$). Після цього для кожного слова i, що входить до тексту, підраховується кількість фрагментів $df(i)$, в яких міститься це слово, а також загальна кількість входження даного слова i у текст – $n(i)$. Після цього розраховується середнє значення TFIDF вагової оцінки для кожного слова за формулою:

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

.

При побудові мереж слів в цій роботі також використовується дисперсійна оцінка ваги слів [17], яка обчислюється наступним чином чином: деяке слово, наприклад A, позначається як A_k^n, де індекс $k = 1,2,...,K$ – номер появи даного слова у тесті, а n – позиція даного слова у тексті. Наприклад, A_3^{50} означає, що на 50-й позиції тексту знаходиться слово A, яке зустрілось третій раз. Інтервалом між послідовними появами слова при таких позначеннях буде величина $\Delta A_k = A_{k+1}^m - A_k^n = m - n$, де на m-й та n-й позиції в тесті знаходиться слово A, яке зустрілось $k+1$-й і k-й рази.

Запропонована в [27] дисперсійна оцінка розраховується як:

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

де: $\left\langle \Delta A \right\rangle$ – середнє значення послідовності $\Delta A_1, \Delta A_2, ..., \Delta A_K$, $\left\langle \Delta A^2 \right\rangle$ – послідовність $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_K^2$, K – кількість появ слова A у тексті.

Ряди з цифрових значень, відповідних словам, перетворяться в графи горизонтальної видимості, в яких вузлам відповідають не лише цифрові значення, але самі слова, що виражають певне змістовне значення. Мережа мови з використанням алгоритму горизонтальної видимості будується в три етапи. На першому на

горизонтальній осі відзначається ряд вузлів, кожен з яких відповідає словам в порядку появи в тексті, а по вертикальній осі відкладаються вагові чисельні оцінки (візуально – набір вертикальних ліній, див. Рис. 1).

На другому етапі будується традиційний граф горизонтальної видимості [21]. Визначається, що між вузлами існує зв’язок, якщо вони знаходяться в “прямій видимості”, тобто якщо їх можна з’єднати горизонтальною лінією, що не перетинає ніяку іншу вертикальну лінію. Цей (геометричний) критерій можна записати, згідно [15,16] таким чином: два вузли (слова), наприклад, B_3^n і C_7^m $(m = n+5)$ поєднуються зв’язком, якщо (рис. 1) $\sigma_n, \sigma_m > \sigma_p$ для усіх $n < p < m$.

Алгоритм побудови можна представити зручним для обчислення способом. Наприклад, на рис. 1 для вузла-слова A_1^{n+2} суміжними в мережі вважаються слова B_3^n та C_1^{n+5} і встановлюються ребра-зв’язки, такі що B_3^n – найближче зліва від A_1^{n+2} слово, з ваговою оцінкою $\sigma_n = \sigma_B$, що перевищує вагову оцінку слова A $\sigma_{n+2} = \sigma_A$, а $C_7\,(m = n+5)$ – найближче з права від A_1^{n+2} слово, для якого $\sigma_{105} > \sigma_{102}$.

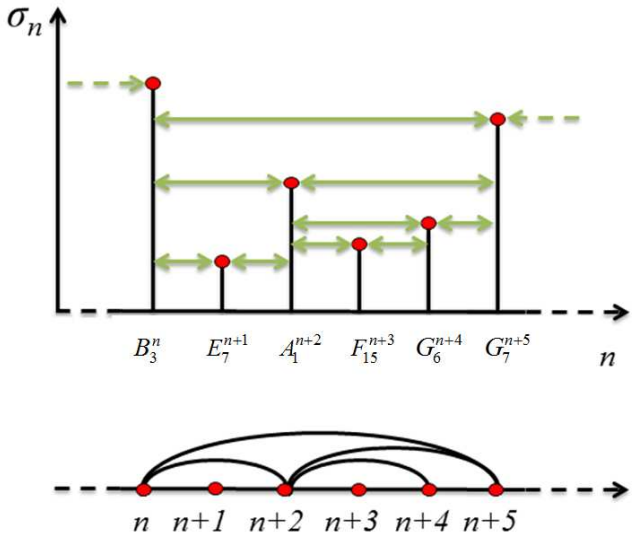

Рис. 1. Приклад побудови графа горизонтальної видимості.

На третьому, завершальному етапі, отримана на попередньому етапі мережа компактифікується. Усі вузли з визначеним словом, наприклад словом A, об’єднуються в один вузол. Усі зв’язки таких вузлів також об’єднуються. Важливо відмітити, що між будь-якими двома вузлами при цьому залишається не більш за один зв’язок – кратні зв’язки вилучаються. Зокрема це означає, що міра (число зв’язків) вузла не перевищує суми степенів $\sum_k A_k^n$. В результаті формується нова мережа мови – компактифікований граф горизонтальної видимості (далі – КГГВ), див. Рис. 2.

Як тексти при побудові мереж мови авторами розглядалася добірка законодавчих актів України, що відносяться до формування та розвитку інформаційного простору держави (Закони України “Про доступ до публічної інформації”, “Про Основні засади розвитку інформаційного суспільства в Україні на 2007 – 2015 роки”, “Про телекомунікації”, “Про захист персональних даних”, “Про основи національної безпеки України”).

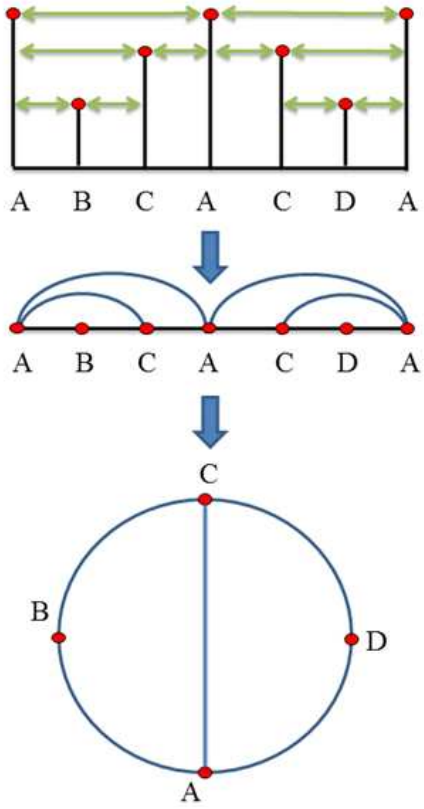

Рис. 2. Етапи побудови компактифікованого графа горизонтальної видимості.

Для усіх побудованих КГГВ-мереж мови було визначено розподіл степенів вузлів, який виявився близьким до статечного ($p(k) = Ck^{\alpha}$), тобто ці мережі є безмасштабними. Були проведені розрахунки параметрів мереж для усіх розглянутих законодавчих актів. У результаті виявилось, що для усіх з них коефіцієнт α змінювався в діапазоні від -1 до $-0{,}95$.

До складу вузлів з найбільшими степенями в КГГВ-мережах, разом із службовими словами (частки, союзи і так далі), потрапили слова, що визначають інформаційну структуру тексту [18, 19].

Для порівняння була досліджена поведінка простих мереж мови, коли на першому етапі побудови мережі встановлюються зв’язки між сусід ними словами, що входять в текст, а на другому – відбувається компактифікація мережі. Очевидно, вага вузлів в цій мережі відповідає частоті появи слів, а їх розподіл – закону Ципфа [20]. При цьому найбільші степені мають вузли, що відповідають словам з найбільшою частотою, що мають велике значення для зв’язності тексту, але мало цікаві для визначення інформаційної структури.

Якщо позначити як Ψ – множину із N різних слів (розглядається випадок $N = 100$), що відповідають найбільш вагомим вузлам наведеної простої мережі мови, а Λ – множину слів, що відповідають найбільш вагомим вузлам КГГВ, то множина $\Omega = \Lambda \setminus \Psi$ відповідає інформативним словам, що мають, крім того, важливе значення і для зв’язності тексту. Авторами досліджувалися результати порівняння 100 найбільш вагомих вузлів для КГГВ-мереж мови за текстами наведених вище законодавчих актів.

Зокрема, в КГГВ-мережі з урахуванням значень TFIDF, по тексту Закону України “Про телекомунікації” до складу множини Ω потрапили такі слова, як “Державне”, “Регулювання”, “Ринку”, “Інтернет”, “Провайдер”, “Трафік”. У КГГВ-мережі для цього

ж тексту за ваговими значеннями слів, відповідними дисперсійним оцінкам, додатково до складу множини Ω потрапили такі слова, як “Суб’єкт”, “Ресурс”, “Переоформлення”, “Рішення”, “Споживачів” та ін.

При аналізі тексту Закону України “Про захист персональних даних” до множини Ω (для КГГВ-мережі з урахуванням вагових значень слів за алгоритмом TFIDF) потрапили такі слова, як “Інформація”, “Відстрочення”, “Орган”, “Баз”, “Виключено”. У КГГВ-мережі для тексту цього законодавчого акту за ваговими значеннями слів, відповідними дисперсійним оцінкам, до складу множини Ω потрапили додатково такіе слова, як “Використання”, “Прав”, “Уповноважений”, “Особа”.

Висновки.

У результаті проведених досліджень:

- Запропоновано алгоритм побудови компактифікованого графа горизонтальної видимості (КГГВ).
- На основі вагових оцінок слів тексту за двома алгоритмами, побудовані КГГВ-мережі мови для текстів різних законодавчих актів.
- Для текстів законодавчих актів серед вузлів, що відповідають КГГВ з найбільшими степенями, присутні слова, не лише важливі для структури тексту, що забезпечують зв’язність, але й ті, що визначають його інформаційну структуру, відбивають семантику тестів.
- Алгоритми визначення ваги слів, що базується на дисперсійній оцінці і TFIDF) виявився близькими за ефективністю на розглянутих прикладах.

Використана література

1. Солганик Г.Я. Синтаксическая стилистика. Сложное синтаксическое целое. – [2-е изд., испр. и доп.] / Г.Я. Солганик. – М. : Высш. шк., 1991. – 182 с.

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

3. Ferrer-i-Cancho R., Sole R.V. The small world of human language // Proc. R. Soc. Lond. – B 268, 2261 (2001).

4. Dorogovtsev S.N., Mendes J. F. F. Language as an evolving word web // Proc. R. Soc. Lond. – B 268, 2603 (2001).

5. Caldeira S.M.G., Petit Lobao T.C., Andrade R.F.S., Neme A., Miranda J.G.V. The network of concepts in written texts // Preprint physics/0508066 (2005).

6. Ferrer-i-Cancho R., Sole R.., Kohler R. Patterns in syntactic dependency networks // Phys. Rev. E 69, 051915 (2004).

7. Ferrer-i-Cancho R. The variation of Zipf's law in human language // Phys. Rev. E 70, 056135 (2005).

8. Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P. Topology of the conceptual network of language // Phys. Rev. E 65, 065102(R) (2002).

9. Sigman M., Cecchi G. A. Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 99, 1742 (2002).

10. Strogatz S. H. Exploring Complex Networks // Nature. – 410. – P. 268-276 (2001).

11. Albert R., Barabasi A.-L. Statistical mechanics of complex networks // Reviews of Modern Physics. – 74. – P. 47 (2002).

12. Nunez A. M., Lacasa L., Gomez J. P., Luque B. Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6. – P. 119-152 (2012).

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

14. Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatoris on words // Physica A. – 390. – P. 2421-2428 (2011).

15. Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. – 28 (1). – P. 11-21 (1972).

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

17. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, – 57(5). – P. 759-764 (2002).

18. Черняховская Л.А. Смысловая структура текста и ее единицы // Вопросы языкознания. – № 6. – С. 118-126. (1983).

19. Giora R. Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam. – 3. – № 2. – P. 155-181 (1983).

20. Zipf G.K. Human Behavior and the Principle of Least Effort. – Cambridge, MA: Addison-Wesley Press. – 573 p. (1949).

21. Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста : *сб. научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика* ; отв. ред. В.Т. Юнгблюд. – Пермь, 2010. – Вып. 1. – С. 85-91.

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