

УДК 004.67

**І. В. Балагура, Д. В. Ланде**

Інститут проблем реєстрації інформації НАН України  
вул. М. Шпака, 2, 03113 Київ, Україна

## **Лінгвістичні дослідження взаємозв'язків науковців на основі аналізу реферативної бази даних «Україніка наукова»**

*Запропоновано використовувати методи комп'ютерної лінгвістики для наукометричних досліджень реферативної бази даних «Україніка наукова». Представлено методичку наукометричних досліджень наукового напрямку на прикладі комп'ютерних наук реферативної бази даних «Україніка наукова» на основі методів комп'ютерної лінгвістики, що дозволяє детально дослідити тенденції наукової співпраці, виділити найбільш комунікативних науковців, наукові групи та визначити ключові слова для певних наукових груп.*

**Ключові слова:** комп'ютерна лінгвістика, реферативна БД «Україніка наукова», мережа співавторства, мережа термінів, наукова взаємодія.

### **Вступ**

Складність визначення кількісного показника ефективності наукових досліджень і необхідність аналізу структури науки та перспектив розвитку науки викликає необхідність дослідження методів наукометрії та розробки відповідних інформаційних систем. Найбільш об'єктивні та визнані інформаційні системи наукометричного аналізу, такі як реферативні бази даних (БД) «Scopus», «Web of knowledge», формуються на основі реферативної інформації та наукометричного інструментарію [1, 2]. Тому актуальною задачею є формування та розвиток української реферативної БД. На сьогодні український реферативний журнал «Джерело» та реферативна база даних «Україніка наукова» є найповнішим джерелом рефератів українських монографій, статей зі збірників наукових праць та серійних (періодичних) видань, матеріалів конференцій, посібників для вузів, авторефератів дисертацій, препринтів [3]. Реферативна БД «Україніка наукова» містить значну кількість української наукової інформації, що є якісною основою для проведення наукометричних досліджень. Наукометричні дослідження реферативної БД «Україніка наукова» проводять науковці з Інституту проблем реєстрації інформації, Національної бібліотеки ім. В.І. Вернадського, Центру досліджень науково-

© І. В. Балагура, Д. В. Ланде

технічного потенціалу та історії науки ім. Г.М. Доброва Національної академії наук України та інших наукових установ [4–7].

Більшість наукометричних інформаційних систем включають методи та технології на основі цитування. Неможливість застосування подібних технологій пояснюється відсутністю обробки посилань в українській реферативній БД. Натомість велика ретроспектива та об'єм даних надають можливість застосування іншого широко розповсюдженого напрямку — методів комп'ютерної лінгвістики.

У попередніх дослідженнях розглянуто оцінку повноти реферативного ресурсу, доцільність використання індексу рубрикатора НБУВ та проблему співпадіння прізвищ та ініціалів авторів [8]. Визначено ряд публікацій авторів, у яких був указаний помилковий або пропущений індекс рубрикатора, що негативно впливає на результати наукометричних досліджень. Визначено прізвища, що є найбільш поширеними в реферативній БД. Проблема ідентифікації авторів також вносить значний вплив на достовірність результатів аналізу. Певні недоліки та особливості реферативної бази даних можна врахувати за допомогою застосування методів комп'ютерної лінгвістики для оцінки наукової галузі.

Комп'ютерна лінгвістика — напрямок у прикладній лінгвістиці, що орієнтується на використанні комп'ютерних інструментів — програм, комп'ютерних технологій організації та обробки даних — для моделювання функціонування мови за тих чи інших умов, ситуацій, проблемних сфер тощо, а також вся галузь застосування комп'ютерних моделей мови в лінгвістиці та суміжних дисциплінах [9]. Необхідно також зазначити розширення можливостей комп'ютерної лінгвістики, що відбуваються завдяки розвитку комп'ютерних наук у цілому: головним об'єктом дослідження стали інформаційні об'єкти, комп'ютерна лінгвістика стала максимально міждисциплінарною, входить до експериментальної парадигми, найчастіше використовує методи математичного моделювання, теорії складних систем і психофізіології (обробки інформації у людини), в комп'ютерній лінгвістиці з'явилися нові об'єкти вивчення (колекції, комп'ютерні кластери і т.ін.) та нові експериментальні можливості (можливості сучасних інформаційних технологій) [9].

Метою досліджень є розробка методики наукометричних досліджень наукового напрямку реферативної бази даних «Україніка наукова» за допомогою методів комп'ютерної лінгвістики.

## **Методика досліджень**

З початку 2000-х років у комп'ютерній лінгвістиці почали застосовувати концепцію складних мереж як побудову мереж мови, тобто мережевої моделі текстових документів [10].

Аналіз мереж — один із потужних методів в інформатиці, що використовується для аналізу знань [11]. Такий метод використовують майже в усіх галузях знань, і кожне застосування містить окремі особливості та потребує розробки власних методик. Мережі співавторів містять зв'язки спільних робіт окремих науковців і характеризують наукову взаємодію. Мережі термінів основуються на статистиці взаємної зустрічаємості ключових термінів в одному документі. Мережі термінів застосовують для опису тематики різних наукових галузей та визначення

трендів розвитку наукових напрямів [11]. Також дані методики використовують у вебметричних дослідженнях для аналізу веб-сторінок. Зарубіжні автори у бібліометричних дослідженнях використовують аналіз термінів, мереж співавторів і цитування комплексно для всебічної оцінки галузей [12].

Методика наукометричного аналізу реферативної бази даних «Україніка наукова» в своїй основі містить 4 етапи, що включають методи складних мереж, методи фільтрації тексту, візуалізацію даних за допомогою програмних засобів Gephi.

На першому етапі визначається галузь і наукові напрямки, за якими буде проведено аналіз, завантажується та фільтрується файл з реферативною інформацією. Результатом першого етапу маємо відфільтровані за певними індексами рубрикуатора дані про авторів і зв'язки між ними, тобто матрицю мережі.

Другим етапом є створення мережі співавторів досліджуваної галузі, а також визначення основних характеристик мережі за допомогою програмних засобів Gephi, а також розрахунок додаткових параметрів за допомогою власних програмних засобів. У результаті другого етапу визначаються основні властивості співробітництва науковців, наукові групи та найбільш комунікативні науковці за певним науковим напрямком.

Третій етап присвячений відбору в повнотекстових базах даних наукових публікацій найбільш комунікативних науковців і створення текстового корпусу для виділення основних термінів за науковими напрямками.

На четвертому етапі проводиться візуалізація мереж термінів науковців і галузі в цілому, розрахунок основних параметрів і виділення так званих опорних слів і відповідних словосполучень. Проводиться узагальнення результатів, опис основних характеристик, тенденцій у галузі.

Розглянемо методику наукометричних досліджень за допомогою методів комп'ютерної лінгвістики на прикладі аналізу комп'ютерних наук в реферативній БД «Україніка наукова».

Перший етап. Для проведення наукометричного аналізу галузі «Комп'ютерних наук», на початковому етапі необхідно визначити відповідні індекси рубрикуатора НБУВ, після чого буде сформовано таблиці вузлів і ребер мережі співавторів. У дослідженні використано рубрики розділів «Інформаційна та обчислювальна техніка» та «Кібернетика»: «Основи інформатики та обчислювальної техніки», «Аналогова і гібридна обчислювальна техніка», «Цифрова обчислювальна техніка», «Комп'ютери і програмування», «Кібернетичні моделі», «Теорія інформації», «Системний аналіз», «Теорія автоматів» та «Біоніка». Обробку даних рубрик і створення мережі співавторів з комп'ютерних наук детально представлено в роботі [13].

Дані з реферативної БД «Україніка наукова» представлені у форматі UKPMARK, що є цифровим форматом представлення бібліографічних даних та українською версією міжнародного комунікативного формату UNIMARK. На мові високого рівня Python було створено програмні засоби, що виконували фільтрацію співавторів за індексами рубрикуатора НБУВ «397. Інформаційна та обчислювальна техніка», «381 Кібернетика» та формували мережу співавторів. Реферативна БД «Україніка наукова» містить записи українською, англійською та російською.

ською мовами, тому інформація про публікації авторів може містити похибки. Для корекції результатів було проведено незначну лінгвістичну обробку прізвищ. Ваги зв'язків між авторами для кожної публікації визначалися обернено пропорційно до кількості її авторів [14].

Другий етап. На цьому етапі формується мережа співавторів та обчислюються її основні характеристики та параметри. За необхідності детального визначення причинно-наслідкових особливостей існує можливість вивчення та візуалізації окремих зв'язків і підмереж. Візуалізацію мережі та обчислення основних характеристик, визначення наукових груп було виконано за допомогою засобів програмного продукту Gephi [15]. У результаті отримано мережу співавторів, що містить 1189 автори та 18049 зв'язки. Діаметр графа, тобто найбільша відстань між двома співавторами через інших авторів, дорівнює 23. Причому середня довжина шляху між вузлами близька до 8. Середня кількість людей, з якими співпрацює один співавтор, або середня степінь вузла, приблизно становить 4. Середня зважена степінь приблизно дорівнює 2, тобто у співавторстві один автор має 2 публікації.

Мережу було поділено на групи на основі розрахунку модулярності вузлів. Модулярність вузла — це величина, що оцінює щільність зв'язків у зв'язній компоненті порівняно зі зв'язками між компонентами. Наявність наукових груп, що можуть мати ознаки наукових шкіл, у галузі «Комп'ютерних наук» можна простежити наочно за особливістю більшості клік, що в своїй основі містять потужного автора з великою кількістю статей у співавторстві та значну кількість маленьких вузлів — авторів-учнів. На рис. 1 наведено фрагмент побудованої мережі співавторів, що містить дані по академіку Палагіну О.В., його співавторам і співавторам його співавторів. Розмір відповідного вузла автора відповідає його кількості зв'язків. Концентрація зв'язків співавторства навколо одного або декількох лідерів можуть характеризувати створення окремих наукових шкіл. У даному випадку їхніми потенційними центрами є такі автори як Палагін О., Сергієнко І., Бодяньський Е., Баркалов А. та інші [14].

Визначення важливих вузлів у мережі є актуальною задачею та потребує детального вивчення предмету дослідження, адже існує багато коефіцієнтів, що надають різнобічні характеристики вершин, та доцільність їхнього застосування визначається тільки відповідністю цілям експериментів. У результаті дослідження було визначено основні показники центральності мережі з комп'ютерних наук [16]. Зважений рівень центральності, що фактично є показником кількості статей у співпраці, відображає об'єм напрацювань автора, а кількість зв'язків автора характеризують коло його співавторів. На рис. 1 рівню центральності вузлів відповідає їхня величина. Автори з високим рівнем показника центральності пов'язані в окремих групах — наукових групах і не пов'язані між собою. В роботі [16] також представлено один із модифікованих показників центральності, що можна застосовувати для визначення потенційних центрів в мережі співавторів.

Третій етап. У дослідженні використовувалися публікації п'яти авторів з комп'ютерних наук російською мовою за останні 10 років, що представлені в повнотекстовій науковій БД «Наукова періодика України» [17]. Для кожного автора було сформовано текстовий корпус, що містив від 15 до 50 статей у фахових журналах, тез доповідей на конференціях і патентів. При попередній обробці тексту було вилучено нетекстові символи та виконано стеммінг.



зв'язки першого (якщо два терміни утворюють третій) та другого (якщо термін утворений двома іншими) роду [19].

**Четвертий етап.** Відібрані терміни використовуються для створення мережі термінів, де вузлами є терміни та словосполучення, а зв'язки відповідають входженню термінів до словосполучень. Такі мережі мови ще називають *L*-простором [10]. На рис. 2 представлено приклад мережі термінів для автора академіка Палагіна О.В., і яка створена за допомогою програмних засобів Gephi.

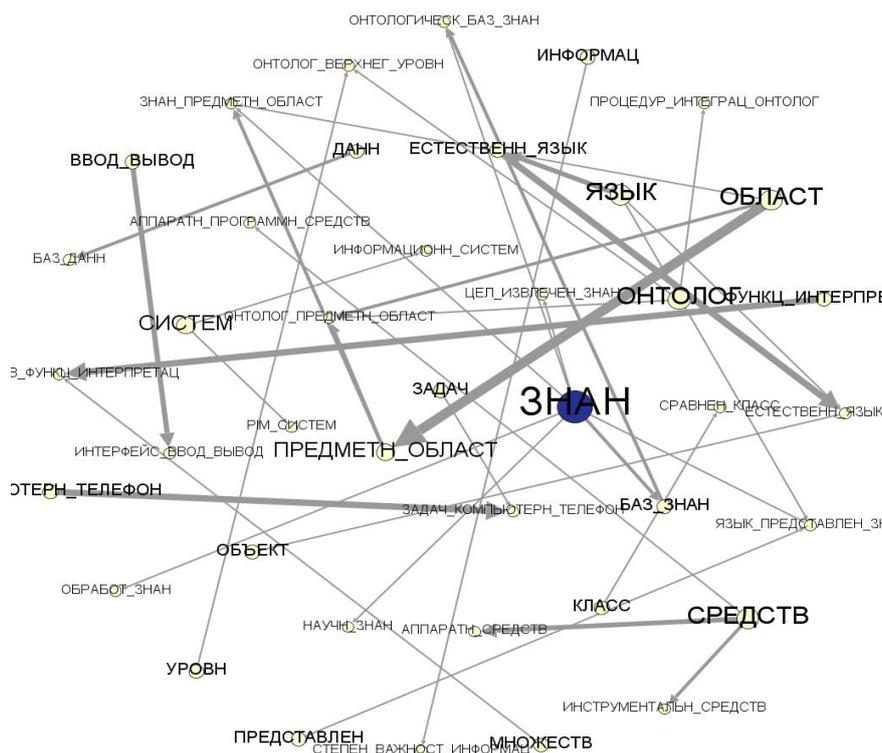


Рис. 2. Мережа термінів на основі аналізу публікацій Палагіна О.В.

Найбільш значимі слова та словосполучення в мережі термінів можна виділяти за допомогою алгоритму HITS (Hyperlink Induced Topic Search — індукований пошук важливих вузлів на основі гіперпосилань). Алгоритм HITS базується на визначенні «авторитетних вузлів» (вузлів, на які багато посилань) та «вузлів-посередників» (вузли, що мають найбільшу кількість посилань). Для автора Палагіна О.В. найбільш значимими термінами є: 1) за значенням «авторитетність» триграми — «онтологія предметної області», «знання про предметну область», «мова представлення знань», «естественно-языковий об'єкт», «онтологія верхнього рівня», «онтологічна база знань», «задачі комп'ютерної телефонії», «множество функцій інтерпретації»; 2) за значенням «посередництво» — «система», «онтологія», «знання», «база знань», «клас», «данні», «задача», «множество», «область», «предметна область», «представлення», «інформація», «об'єкт», «мова» та ін. Таким чином, можна виділити серед основних інтересів автора методи комп'ютерної лінгвістики, бази знань.



## Висновки

Запропоновано використовувати методи комп'ютерної лінгвістики для наукометричних досліджень взаємозв'язків науковців на основі аналізу реферативної бази даних «Україніка наукова». Запропоновано методику наукометричних досліджень наукового напрямку реферативної бази даних «Україніка наукова» на основі методів комп'ютерної лінгвістики, що дозволяє детально дослідити тенденції наукової співпраці, виділити найбільш комунікативних науковців, наукові групи та визначити ключові слова для певних наукових груп. Показано можливість застосування методики наукометричного аналізу на прикладі аналізу комп'ютерних наук у реферативній базі даних «Україніка наукова». Отримано опис основних характеристик, тенденцій у галузі комп'ютерних наук, виділено основні наукові групи та найбільш комунікативних науковців, основні терміни, створено візуалізацію мереж термінів.

1. *Scopus* [Електронний ресурс]. — Режим доступу: <http://www.scopus.com>. — Назва з екрану.
2. *Web of Knowledge* [Електронний ресурс]. — Thomson Reuters, 2012 — Режим доступу: <http://wokinfo.com>. — Назва з екрану.
3. *Реферативна база даних «Україніка наукова»* [Електронний ресурс]. — Режим доступу: <http://nbuv.gov.ua/db/ref.html>. — Назва з екрану.
4. *Значення видання українського реферативного журналу «Джерело» для розвитку наукових комунікацій в Україні* / А.А. Крючин, Л.Й. Костенко, Н.М. Мініна [та ін.] // *Наука України у світовому інформаційному просторі*. — 2012. — Вип. 6. — С. 20–23.
5. *Наукова періодика України та бібліометричні дослідження: [монографія]* / [Л.Й. Костенко, О.І. Жабін, С.О. Копанєва, Т.В. Симоненко]. — НАН України, Нац. б-ка України ім. В.І. Вернадського. — К., 2014. — 173 с.
6. *Рибачук В.П.* Методологічні проблеми оцінювання продуктивності наукової діяльності / В.П. Рибачук // *Наука та наукознавство*. — 2013. — № 2 (80). — С. 46–52.
7. *Букшина Т.Ф.* Бібліометричний аналіз наукових публікацій з питань дошкільної освіти і виховання в загальнодержавній реферативній БД «Україніка наукова» (1998–2012 рр.) [Електронний ресурс] / Т.Ф. Букшина // *Інформаційні технології і засоби навчання*. — 2014. — Т. 40, Вип. 2. — С. 135–150. — Режим доступу: [http://nbuv.gov.ua/j-pdf/ITZN\\_2014\\_40\\_2\\_15.pdf](http://nbuv.gov.ua/j-pdf/ITZN_2014_40_2_15.pdf)
8. *Балагура І.В.* Дослідження мережі співавторів з бібліотечної справи та наукознавства [Електронний ресурс] / І.В. Балагура // *Міжнар. наук. конф. «Місце і роль бібліотек у формуванні національного інформаційного простору»*. — Жовтень, 2014. — м. Київ. — Режим доступу: <http://conference.nbuv.gov.ua/report/view/id/363>. — Назва з екрану.
9. *Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие* / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ [и др.]. — М.: МИЭМ, 2011. — 272 с.
10. *Ланде Д.В.* Елементи комп'ютерної лінгвістики в правовій інформатиці / Д.В. Ланде. — К.: НДПП НАПрН України, 2014. — 168 с.
11. *Mining term networks from text collections for crime investigation* / Yuen-Hsien Tseng, Zih-Ping Ho, Kai-Sheng Yang, Chun-Cheng Chen // *Expert Systems with Applications*. — 2012. — Vol. 39, Issue 11. — P. 10082–10090.

12. *Chappin E.J.L.* Transition and transformation: A bibliometric analysis of two scientific networks researching socio-technical change / Emile J.L. Chappin, Andreas Ligtoet // *Renewable and Sustainable Energy Reviews*. — 2014. — Vol. 30. — P. 715–723.
13. *Ланде Д.В.* Наукометричні дослідження мереж співавторства по базі даних «Україніка наукова» / Д.В. Ланде, І.В. Балагура // *Реєстрація, зберігання і оброб. даних*. — 2012. — Т. 14, № 4. — С. 41–51.
14. *Визначення* потенційних експертних груп науковців у мережі співавторства з вико ристанням методів підтримки прийняття рішень / І.В. Горбов, С.В. Каденко, І.В. Балагура [та ін.] // *Реєстрація, зберігання і оброб. даних*. — 2013. — Т. 15, № 4. — С. 87–97.
15. *The Open Graph Viz Platform Gephi* [Електронний ресурс]. — Режим доступу: <https://gephi.github.io/>. — Назва з екрану.
16. *Балагура І.В.* Характеристики сети соавторов медицинских наук / И.В. Балагура, Д.В. Ландэ, И.В. Горбов // *Клин. информат. и телемед.* — 2013. — Т. 9, № 10. — С. 141–144.
17. *Наукова періодика України* [Електронний ресурс]. — Режим доступу: [http://www.irbis-nbuv.gov.ua/cgi-bin/irbis\\_nbuv/cgiirbis\\_64.exe?C21COM=F&I21DBN=UJRN&P21DBN=UJRN&S21CNR=20](http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?C21COM=F&I21DBN=UJRN&P21DBN=UJRN&S21CNR=20). — Назва з екрану.
18. *Ландэ Д.В.* Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текстов на различных языках [Електронний ресурс] / Д.В. Ландэ, А.А. Снарский, Е.В. Ягунова // *Тези доповідей Міжнар. наук. конф. «Горизонти прикладної лінгвістики і лінгвістичних технологій» («Megaling-2013»)*. — Київ, 2013. — Режим доступу: <http://megaling.ulif.org.ua/prezentatsiyi-dopovidey/>. — Назва з екрану.
19. *Ландэ Д.В.* Подход к созданию терминологических онтологий / Д.В. Ландэ, А.А. Снарский // *Онтология проектирования*. — 2014. — Т. 12, № 2. — С. 83–92.
20. *Ягунова Е.В.* Эксперимент и вычисления в анализе ключевых слов художественного текста: сб. науч. трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика. — Пермь, 2010. — Вып. 1. — С. 85–91.

Надійшла до редакції 09.09.2014